Stylistics *versus* Statistics:

A corpus linguistic approach to combining techniques in forensic authorship analysis using Enron emails

David Wright

Submitted in accordance with the requirements for the degree of Doctor of Philosophy

The University of Leeds

School of English

September 2014

The candidate confirms that the work submitted is his/her own, except where work which has formed part of jointly authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

The work in Chapter 6 of the thesis is an expansion of the experiment performed and reported in:

Johnson, Alison and David Wright. 2014. Identifying idiolect in forensic authorship attribution: an n-gram textbite approach. *Language and Law (Linguagem e Direito)* 1(1), 37–69.

In the above article, I was responsible for the experiment and the following sections are directly attributable to me:

‘Computational tools’ (p. 44)
‘Jaccard’s similarity coefficient’ (p. 44–5)
‘Experimental set up’ (p. 45–6)
‘Attribution task’ (p. 56–7)
‘Results’ (p. 57–60)

In the above article, the following sections are directly attributable to Dr Alison Johnson:

‘James Derrick’ (p. 43–4)
‘Case study of James Derrick’ (p. 46)
‘Identifying Derrick’s professional authorial style’ (p. 46–54)
‘Conclusion’ (p. 62–63)

The remaining sections of the above article are co-authored:

‘Introduction’ (p. 38–40)
‘N-grams, idiolect and email’ (p. 40–42)
‘The Enron Email Corpus’ (p. 42–43)
‘Summary of findings’ (p. 54–56)
‘Derrick’s discriminating n-grams’ (p. 60–62)

Acknowledgements

First and foremost, my deepest thanks go to my wonderful parents Janet and Joseph William Wright, for their unwavering support throughout my PhD and beyond. Their selflessness is such that when the idea of me doing a PhD first arose, they offered to sell the family home to finance it. Thankfully, that was not necessary (I would never have allowed them to sell it, anyway). They have had to maintain me during the times when I worked from home in the North East, though. I hope that they know that I love them very much.

I would also like to thank the rest of my close family and friends, who have been consistently supportive during my seven years at University. Special mention must go to someone I have known and loved for the last three of those years: my girlfriend Rachael Groenendaal. She has known me only as a PhD student, so I am looking forward to sharing with her my less anxious, nervous and restless side. The patience and understanding she has is immeasurable, and every day I am reminded that I am lucky to have such an extraordinary person in my life.

This PhD would not have been possible without the exceptional supervision of Dr Alison Johnson. Alison has the unique ability to calm me when I am worried, and inspire me when I am doubtful. It is that ability, combined with her remarkable trust in me, which has guided me through this work. I hope that one day, I can mentor someone in the way Alison has mentored me.

Special thanks also go to David Woolls for enduring countless ridiculous questions and requests from me. David has responded every time without complaint, and his software and insightful comment have been invaluable to this research. It was him, along with Alison, who first introduced me to the Enron corpus.

Finally, I would like to extend my thanks to past and present lecturing staff in the School of English at the University of Leeds, in particular: Clive Upton, Anthea Fraser Gupta and Fiona Douglas. Without their inspirational teaching and critical feedback during my time as an undergraduate and postgraduate, I would not be pursuing an academic career.

This study is funded by an Arts and Humanities Research Council Doctoral Studentship (Grant number AH/J502686/1).

Abstract

This thesis empirically investigates how a corpus linguistic approach can address the main theoretical and methodological challenges facing the field of forensic authorship analysis. Linguists approach the problem of questioned authorship from the theoretical position that each person has their own distinctive idiolect (Coulthard 2004: 431). However, the notion of idiolect has come under scrutiny in forensic linguistics over recent years for being too abstract to be of practical use (Grant 2010; Turell 2010). At the same time, two competing methodologies have developed in authorship analysis. On the one hand, there are qualitative stylistic approaches, and on the other there are statistical 'stylometric' techniques. This study uses a corpus of over 60,000 emails and 2.5 million words written by 176 employees of the former American company Enron to tackle these issues in the contexts of both authorship attribution (identifying authors using linguistic evidence) and author profiling (predicting authors' social characteristics using linguistic evidence).

Analyses reveal that even in shared communicative contexts, and when using very common lexical items, individual Enron employees produce distinctive collocation patterns and lexical co-selections. In turn, these idiolectal elements of linguistic output can be captured and quantified by word n-grams (strings of *n* words). An attribution experiment is performed using word n-grams to identify the authors of anonymised email samples. Results of the experiment are encouraging, and it is argued that the approach developed here offers a means by which stylistic and statistical techniques can complement each other. Finally, quantitative and qualitative analyses are combined in the sociolinguistic profiling of Enron employees by gender and occupation. Current author profiling research is exclusively statistical in nature. However, the findings here demonstrate that when statistical results are augmented by qualitative evidence, the complex relationship between language use and author identity can be more accurately observed.

Table of Contents

List of Tables

List of Figures

1 Introduction

> This sally into the relatively uncultivated field of 'forensic linguistics' has been interesting for a number of reasons, but two in particular. Firstly, it has provided the linguist with one of those rare opportunities of making a contribution that might be directly useful to society…
>
> (Svartvik 1968: preface)

The above quotation contains the first published use of the term 'forensic linguistics', which is defined today as 'the application of linguistics to three principle domains: written legal texts, spoken legal practices and the provision of evidence for criminal and civil investigations' (Coulthard et al. 2011: 529). This thesis is situated within the third of these applications, the use of language as evidence. In particular, this research is of benefit to evidential work in cases of forensic authorship analysis, which are concerned with answering questions of 'who wrote this text?' (authorship attribution) and 'what kind of person wrote this text?' (author profiling). Authorship attribution is the process in which a linguist attempts to identify the author of an anonymous or 'disputed' document based on the linguistic clues left behind by the writer. In a forensic context, the documents in question are potentially evidential in alleged infringements of the law or threats to security. Author profiling also draws on clues left behind by writers, but rather than attempting to attribute a text to a specific individual, the task is to predict as much as possible about what kind of person wrote it, in terms of social characteristics such as age, gender, level of education, native language and personality.

Unfortunately, in today's digital world, the skills of forensic linguists are required far more commonly than Svartvik's statement would suggest. For example, in June 2005 teenager Jenny Nicholl was murdered by her lover David Hodgson. Linguist Malcolm Coulthard was consulted to make a judgement on whether a series of text messages sent from her mobile telephone after her disappearance were written by her or Hodgson. Coulthard compared a range of linguistic features exhibited in the disputed text messages with those in 'known' (that is, known to be written by) text messages written by both Nicholl and Hodgson. He arrived at the opinion that Nicholl was unlikely to have written the messages, while Hodgson was one of a small group of potential authors. Hodgson was convicted of murder in

February 2008 (*BBC News* 2008). More recently, in 2013, Tim Grant and Jack Grieve from Aston University's Centre for Forensic Linguistics were consulted by police in a case in which a husband, Jamie Starbuck, was suspected of murdering his wife Debbie Starbuck (*The Telegraph* 2013a). Jamie and Debbie were travelling the world together and Debbie's family regularly received emails thought to be sent from the couple. However, once Debbie's family became suspicious that her writing style had changed, Grant and Grieve were contacted by the police to investigate whether the emails were being authored by Debbie or her new husband. Using a similar approach to Coulthard, and comparing the disputed emails against known emails from both Debbie and Jamie, Grant and Grieve were able to pinpoint the exact date at which Debbie's email style shifted to become more similar to her husband's. This date matched that on which the police believed Debbie had been killed, and Jamie was arrested, convicted and sentenced to life imprisonment in May 2013. Authorship attribution made international headlines later in 2013, but this time the law had not been broken. Using computational techniques, and following a tip-off from *The Sunday Times,* Patrick Juola was able to identify J.K. Rowling as the author of *The Cuckoo's Calling,* which she had written and published under the pseudonym 'Robert Galbraith' (*The Telegraph* 2013b).

All of these high profile authorship analysis cases were preceded by Coulthard's (1994) report on the infamous case of Derek Bentley. In 1953, at the age of nineteen, Derek Bentley was hanged for the murder of a police officer, Constable Sidney Miles, during an attempted burglary. Although Bentley's accomplice, Christopher Craig, fired the fatal shot, Bentley was convicted and sentenced as party to the crime. Central to Bentley's conviction was the incriminating evidence included in the statement he gave to the police following his arrest. Bentley was illiterate, and claimed in court that the police officers had 'helped' him with his statement, a claim which the officers denied on oath. In his analysis of the disputed statement, Coulthard focused on a range of linguistic features, one of which was the word *then*. Using the spoken data from the Bank of English corpus, and two comparable corpora of legitimate witness statements and police statements, Coulthard concluded that the frequency of use of the word *then* was unusually high in the statement purporting to be written by Bentley, and that its frequency of use, and its grammatical placement after the clause's subject, was more consistent with the register of police writers. In 1998, 45 years after Bentley's hanging, he was

granted a posthumous pardon. Coulthard (1994: 40) concluded that improved methodologies for authorship analysis 'must depend, to a large extent, on the setting up and analysing of corpora'.

However, over the last twenty years corpus linguistic approaches to authorship analysis have not developed. Despite intermittent reminders of the value of such methodologies (e.g. Kredens 2002; Tiersma and Solan 2002; Cotterill 2010; Kredens and Coulthard 2012) the field has been taken in quite a different direction. The current situation in authorship analysis is that there are two major methodological approaches to analysing and attributing authorship of disputed documents. At best, these two approaches are divergent, and at worst, they are competing. On the one hand, there are traditional 'stylistic' approaches, relying on the manual analysis of linguistic variation, consistency and distinctiveness across texts, such as those utilised by Coulthard, Grant and Grieve in the cases above. Despite having been relied upon as expert testimony in court in such cases, critics of this approach argue that it is too heavily reliant on the subjective intuitions of the experts involved, and so any results produced may be considered unreliable. On the other hand, there are 'stylometric', algorithmic and automated approaches to distinguishing between and categorising texts and authors on the basis of the relative frequencies of pre-defined linguistic features and various statistical techniques. Although these stylometric studies refer to the datasets they use as 'corpora', the methodologies they apply cannot be described as corpus linguistic, at least not insofar as McEnery and Hardie's (2012: 1) definition of corpus linguistics as focusing 'upon a set of procedures, or methods, for studying *language'* (my emphasis). Although such studies provide seemingly accurate and reliable results in identifying authors, little evidence is offered in terms of the precise nature of the linguistic similarity between the texts involved, making these results difficult to communicate to courts and juries.

Alongside these developments, the field of author profiling has emerged, and with research being produced exclusively by computational linguists, the approaches to this area of authorship analysis are all stylometric in nature. There is a potential risk, however, that should this research trend continue, then authorship profiling and its practitioners face the prospect of finding themselves in a methodological dilemma akin to that which authorship attribution is currently enduring. Specifically,

the procedures they use and advocate may provide impressive statistical results, but there is little linguistic evidence to support them.

In addition to the methodological issues posed by two apparently incompatible schools of thought, a theoretical debate has begun surrounding the concept on which authorship analysis is founded: that each individual has their own distinctive idiolect (Coulthard 2004: 431). Despite being one of the most familiar and well-established theories in linguistics, there is little empirical evidence to support the existence of distinctive idiolects. This has implications for authorship analysis, and over recent years forensic linguists (Grant 2010; Turell 2010) have worked to reconceptualise what 'idiolect' means in practical terms. Emphasis has been shifted away from the abstract notion of 'idiolect' being everything an author *could* say at any given moment, towards focusing on the linguistic features the author uses in their writing, and measuring how common or rare these features are in the population of writers from which the author is taken.

Using a dataset of 60,000 emails and 2.5 million words written by employees of the former American energy company Enron, the central aim of this study is to investigate how a corpus linguistic approach can be used to address both the methodological and theoretical challenges that the field of authorship analysis currently faces. Its title, '*Stylistics* versus *Statistics*' serves not to place these approaches in opposition. Rather, on the contrary, the aim here is to investigate how corpus linguistics can offer a means by which qualitative and quantitative techniques can be used in combination in authorship analysis, as it has offered various other fields of linguistic enquiry for decades.

1.1 Thesis outline

Chapter 2 of this thesis gives a comprehensive overview of the theoretical and methodological challenges in modern day authorship analysis introduced above. First, the changing nature of the theory of idiolect is documented, from its inception in the nineteenth century, to its current status in forensic linguistics. Following this, various stylistic and stylometric approaches to authorship attribution are described, with a particular focus on how they have a shared interest in relying on lexical variation to distinguish between authors and attributing disputed documents. The respective benefits and drawbacks of the two approaches are then highlighted, and the very recent movements within forensic linguistics to bridge the gap between the

two are discussed. The penultimate section of Chapter 2 draws on existing research and theory from corpus linguistics and psycholinguistics in proposing collocation patterns and word n-grams as a means by which individuals' idiolects can be identified, captured and utilised in a method for authorship analysis. Finally, the early work emerging from the field of author profiling is reviewed, and compared with the more established profiling procedures in forensic psychology.

Chapter 3 introduces the Enron Email Corpus. After a brief description of the company and its demise, attention is shifted to the dataset, the various forms in which it is available, and the areas of research which have made use of it. Following this, the procedure which was undertaken in the extraction and preparation of data for this study is detailed and the ethical considerations are outlined. Then the corpus itself is described and evaluated in terms of its usefulness for authorship analysis. Finally, the different samples which are drawn from the corpus and utilised in the various analysis chapters of the thesis are presented.

Chapter 4 details the computational tools and procedures that are used in the analysis of the Enron corpus, which include existing commercially available software packages, and a bespoke piece of lexical analysis software developed especially for this study. Following this, the various statistical techniques used throughout the thesis are introduced and explained.

Chapter 5 is the first of three analysis chapters. It sets out to provide empirical evidence in support of idiolect. By using the Enron Email Corpus as a reference set representing the population norms of a linguistic community, the rarity and distinctiveness of collocation patterns and lexical sequences are measured. The analysis focuses specifically on three very common words in the corpus and investigates the extent to which, even within elements of a lexical repertoire shared across the corporation, and within very common linguistic practices, author-distinctive lexical co-selection preferences can be identified.

Building on the theoretical foundations laid in Chapter 5, Chapter 6 proposes an innovative approach to authorship attribution. Drawing on word n-grams as quantifiable linguistic features which capture idiolectal collocation patterns, the aim is to develop a methodology which combines the objective and statistical measurement of author style with theoretically-underpinned and linguistically-explainable results. The proposed method is systematically and rigorously tested in

terms of its accuracy in attributing email samples, of various sizes, to their correct authors.

In the final of the three analysis chapters, Chapter 7 extends the combined statistical and stylistic approaches developed in the previous two chapters to the problem of author profiling. The first part of the analysis continues the developing tradition of quantitatively identifying the relationship between linguistic features and particular social characteristics, specifically the gender and occupation of Enron employees. The second part of the analysis selects three such features, and examines the different ways in which they are used both across the groups and by different members of the same group. Such qualitative evidence, it is argued, provides richer insight into the use of linguistic features and author identity, and can be used to supplement the purely quantitative results.

Chapter 8 bring the thesis to a conclusion. It summarises the major findings and assesses the contributions it makes, both theoretically and methodologically, to the fields of authorship attribution and author profiling. Finally, some suggestions are made as to the directions for future research, both in terms of developing corpus linguistic approaches to authorship analysis and for the field more generally.

2 Theoretical and methodological issues in authorship attribution and profiling

This chapter details the theoretical and methodological challenges that authorship attribution and author profiling are currently facing. First, the theoretical notion of 'idiolect' is discussed, and in particular how it has been developed by forensic linguists. Second, the two divergent methodological approaches to authorship attribution are described. The relative advantages and disadvantages of both are compared, recent moves to combine the two are discussed, and the corpus linguistic approach used in this study to combine them is introduced. Next, the early work in the developing field of author profiling is reviewed, and a suggestion is made with regard to the direction for this young field. Finally, the chapter concludes with an outline as to how the various analyses of this thesis aim to address these theoretical and methodological challenges.

2.1 The usefulness and utility of a theory of idiolect

> 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*.
>
> (Coulthard 2004: 431)

Neogrammarian Hermann Paul (1888) is credited with the first major discussion of the notion of linguistic individuality, as he argued that 'every linguistic creation is always the work of one single individual only' (Paul 1889: xliii). Sapir (1927: 903–4) distinguishes between the social norm and individual expression, stating that 'we all have our individual styles in both conversation and address'. Sapir's contemporary, Bloomfield (1933: 45), also highlighted the uniqueness of individuals' language choices, claiming that 'if we observed closely enough, we should find that no two persons—or rather, perhaps, no one person at different times—spoke exactly alike'. The term 'idiolect', compounded from the Greek *idios* meaning 'one's own' and *lektos* meaning 'chosen expression or word' (Kuhl 2003: 4), was coined by Bloch (1948: 7). He defined it as 'the totality of the possible utterances of one speaker at one time in using language to interact with one other

speaker'. He clarifies that 'an idiolect is not merely what a speaker says at one time: it is everything that he *could* say in a given language' (Bloch 1948: 7, original emphasis). Throughout the twentieth century, definitions of idiolect have remained relatively consistent. Hockett's (1958: 321) definition, for example, was: 'the totality of speech habits of a single person at a given time constitutes an idiolect'. Similarly, Decamp (1969: 18) defines idiolectal grammar as 'a specific finite set of rules of an individual speaker-hearer's linguistic competence'. More recently, Dittmar (1996: 111) describes idiolect as 'the language of the individual, which because of the acquired habits and the stylistic features of the personality differs from that of other individuals'.

There is, then, a clear consensus in linguistics that individuals have their own unique linguistic patterns and preferences. However, as Johnstone (1996: 13–14) describes, 'the individual is often anywhere but in the center of interest' in sociolinguistic research. She identifies Saussure's (1966 [1916]) distinction between *langue* (as the abstract systematic principles of language) and *parole* (events of language use), as a main reason as to why 'quantitative sociolinguists focus on the linguistic system rather than on the individual speaker' (Johnstone 1996: 14). This preoccupation with the group at the expense of the individual is made clear by Labov (1972: 277) who argues that:

> we define language [...] as an instrument used by members of the community to communicate with one another. Idiosyncratic habits are not part of language so conceived, and idiosyncratic changes no more so.

Because idiolectal variation has been traditionally relegated to the periphery of the processes of language variation and change in this way—and that it is impossible to collect the data required to investigate everything a person *could* say at any given time (Bloch 1948: 7)—there has been very little serious sociolinguistic research into the language of individuals. This has resulted in the situation outlined by Kredens (2002: 405) who states that while there is 'universal agreement that an individual's linguistic repertoire is in some way distinct', this claim has 'not thus far been supported by empirical research'. This is echoed by Barlow (2010: 1), who says that with no other theory in linguistics is there 'such a large gap between the familiarity of the concept and lack of empirical data on the phenomenon'.

As Coulthard (2004: 431) in the quotation above outlines, the assumption that individuals have their own distinctive idiolects is the basis on which authorship

attribution is performed. However, as a result of the abstract nature of idiolect—a theoretical construct without empirical evidence to support it—questions have been raised in recent years as to its usefulness in forensic linguistics. Turell (2010: 216–7) acknowledges that 'idiolects can only be determined with countless amounts of data from each individual' and suggests that 'idiolects have been accepted as *a priori* constructs, and their existence assumed too readily'. As an alternative to idiolect, Turell (2010: 217) proposes the notion of 'idiolectal style', which she argues 'could be more relevant to forensic authorship contexts':

> 'Idiolectal style' would have to do primarily, not with what system of language/dialect an individual has, but with a) how this system, shared by lots of people, is used in a distinctive way by a particular individual; b) the speaker/ writer's production, which appears to be 'individual' and 'unique' (Coulthard 2004) and also c) Halliday's (1989) proposal of 'options' and 'selections' from these options.

With this notion, Turell is moving away from the abstract concept of idiolect as all possible linguistic production by an individual, towards actual linguistic output as manifest in their speech or writing. Grant (2013: 473) goes further, as he asserts that 'the idea that comparative authorship analysis rests upon a strong theoretical assertion of an idiolect is false', and that 'the empirical discovery of [linguistic] consistency and distinctiveness' can be a sufficient foundation for authorship work. Like Turell, Grant's focus is on linguistic output as exhibited in the writing of individuals, as opposed to being concerned with the abstract and idealised concept of idiolect. Grant (2010: 515; 2013: 474) draws a clear difference between the type of 'distinctiveness' required for reliable authorship attribution and the type of 'distinctiveness' which has more profound implications for theoretical discussions of idiolect. When discussing the practicalities of an authorship case, the distinctiveness Grant is referring to is 'pairwise-distinctiveness'; he argues that 'it may not be necessary to show a writer's distinctiveness against all possible authors', but rather it 'may only be necessary to compare one author with other relevant authors in the case' (Grant 2013: 474). On the other hand, if one person's linguistic style can be said to be distinctive against a reference population of writers, this is referred to as 'population-level distinctiveness' (Grant 2010: 515). When a person's linguistic style stands out as being unique against a large population of writers, we begin to gain a sense of the idiolectal nature of the author's use of particular

linguistic features and patterns. Such a finding, Grant (2010: 522) argues, would be 'an astounding fact'.

Grant's proposal of comparing an individual's use of a particular linguistic feature or set of features against a population-level knowledge of how these features are used more widely, is the same as Turell's (2010: 217) notion of 'Base Rate Knowledge'. However, Turell (2010: 240) claims that to obtain a Base Rate Knowledge with regard to the rarity or expectancy of a particular linguistic feature or pattern requires 'corpora consisting of all possible existing written idiolects of all writers'. Over more recent years, however, emphasis seems to have shifted away from the strictest all-encompassing sense of population-level or Base Rate Knowledge, to knowledge of a more specific or relevant nature. Nini and Grant (2013: 192–3), for example, use Biber's (1988) multidimensional framework results as a Base Rate Knowledge for the frequency of particular linguistic features across different genres. By comparing authors' use of certain linguistic features against this Base Rate Knowledge offered by Biber's results, they can go some way towards measuring how distinctive their authors' linguistic preferences are against this norm. Similarly, a few years after outlining a daunting, if not impossible, criterion for Base Rate Knowledge, Turell (in Turell and Gavaldà 2013) redefines the concept:

> This Base Rate Knowledge implies the collection of data regarding the general usage of the linguistic parameters being considered *by a relevant population*, or group of language users *from the same linguistic community*, with which the specific behaviour of the speakers or writers under comparison can be compared.
>
> (Turell and Gavaldà 2013: 499, my emphasis)

Here, Turell and Gavaldà are emphasising the importance of Base Rate Knowledge about a particular linguistic feature or pattern being taken from a 'relevant population', and from writers within the 'same linguistic community' as the author and text in question. This kind of specificity has replaced the requirement for immeasurable (and uncollectable) reference data from all of the writers and speakers of a given language. This reflects earlier calls for such relevant reference data in forensic authorship attribution. Kredens and Coulthard (2012: 507, 512), for example, argue that 'the collection of specialised corpora to provide population-specific statistics on usage' is 'one way forward for forensic linguistics'. Furthermore, Kredens (2002: 435) argues that a welcome development would be

'the development of reference corpora, serving as sources of normative data'. However, he qualifies this by explaining that any reliable reference data should be characterised by biological, social and interactional variables identical with those of the questioned documents against which they are being compared. Grant (2013: 473) also emphasises the importance of population data being 'relevant':

> It must be recognized, however, that the greater the degree of consistency in any comparison corpus, the greater the weight of evidence there will be for an attribution. Identifying consistency within relevant texts also requires the creation of a linguistically *relevant* comparison corpus, which accounts for genre as well as other sources of linguistic variation.

Forensic linguistics, and in particular authorship analysis, stands to gain the most from a usable theory of idiolect. Through the introduction and discussion of notions such as population-level distinctiveness and Base Rate Knowledge, unique individual language variation is no longer the idealised concept first proposed by Paul, Sapir, Bloomfield and Bloch. Instead, it is becoming an empirically analysable phenomenon. Given suitable population data representing the norms of the speech community from which a writer is taken, serious claims can be made about the idiolectal nature and uniqueness of linguistic features and patterns found in their writing. One of the main motivations of corpus linguistics since its inception has been to empirically test linguistic theory (McEnery and Wilson 2001: 5; McEnery and Hardie 2012: 168). The Enron Email Corpus offers normative reference data of a relevant population of writers from within the same linguistic community (see Section 3.4.1). The main aim of Chapter 5 of this thesis is to use this population-level data to produce empirical, robust and reliable linguistic evidence in support of the theory of idiolect as manifest in the writing of individual Enron employees. In turn, the value and implications, both theoretical and methodological, for forensic authorship analysis are discussed.

2.2 Approaches to identifying authorship

> Already there have developed a variety of different approaches in forensic case work and these are often considered as being in competition with one another.
>
> (Grant 2008: 225)

The current situation in authorship analysis is that there are two main methodological approaches to attributing authorship of disputed or anonymous documents. At best, these two approaches are divergent, and at worst, as noted by

Grant above, they are competing. Traditional stylistic approaches which are based on the qualitative examination of texts and identification of linguistic variation are at odds with purely quantitative, computational and 'stylometric' techniques which rely on algorithmic statistics in measuring similarity between texts and authors. What these approaches have in common, however, is that they both predominantly focus on identifying lexical variation between authors, that is, their use of words. Holmes (1994: 90–1) offers a reason for this emphasis on lexis, commenting that 'word-usage offers a great many opportunities for discrimination [between authors]', and drawing on the arguments of Tallentire (1973; 1976) states that 'the lexical level is the obvious place to initiate stylistic investigations, since [...] more data exist at the lexical level than at any other'.

The subsections that follow review the lexical approaches to authorship attribution from both the stylistic and the statistical perspectives, while highlighting the advantages and criticisms of both, before suggesting a complementary corpus-based approach to utilising lexical variation for forensic linguistic purposes.

2.2.1 Lexis in stylistic approaches to authorship attribution

Coulthard's (1994) analysis of Derek Bentley's questioned witness statement detailed in the Introduction (Chapter 1) is arguably the most seminal study in forensic authorship analysis. Another well-reported case is that of the 'Unabomber' in the United States of America (Coulthard 2004: 432; Fitzgerald 2004; Cotterill 2010: 584). Between 1978 and 1995 a series of bombs were sent through the post to universities and airlines across the USA (hence '*Una*bomber'). In 1995, six national newspapers received a 35,000 word document from a writer claiming to be the Unabomber, offering to stop sending bombs if the document was published. Months later, after receiving a tip-off, the FBI arrested a man and seized a series of documents (his known writings) from his home. One of these was a 300-word document, against which the FBI compared the 35,000 word manifesto and found substantial similarities between them in terms of lexical and phrasal choice. In particular, twelve items were identified as being indicative of common authorship, including *thereabouts*, *gotten* and *at any rate*. The defence hired a linguist who argued that no significance could be attached to these lexical similarities, and claimed that these twelve words could be expected to occur in any text that was arguing a case (Coulthard 2004: 433) as this was. In probably the first use of the web as a corpus (Cotterill 2010: 584) the FBI performed an internet search and

found only 69 documents online in which all of these twelve words and phrases occurred. However, each of these 69 documents was an electronic version of the 35,000 word manifesto. This countered the defence claim and was extremely strong evidence of these lexical choices being idiolectal to the arrested man.

Winter (1996) analyses the vocabulary use in three short written confessions, one of which the accused in the case denies writing. He (Winter 1996: 150) compares the most frequent vocabulary and investigates concordance lists showing how these words are used in context within the confessions. Based on these lexical analyses, he concluded that the disputed confession was made up largely of words which do not appear in the known writings of the author, and words which are shared between the known and disputed writings are used in different ways (Winter 1996: 165–6).

Coulthard (2013) reports a case in which there was one single disputed email sent from the account of Mr. Stephen Goggin (a pseudonym) to a man named Mr. Dennis Juola. Given the content of the email, the number of candidate authors of the disputed email was narrowed to four: Mr. Goggin, Mr. Tim Widdowson, Mr. John Shuy, and Ms. Janet Gavalda. Coulthard was given access to a huge amount of comparison material, including a 190,000 email database of the company, as well as sets of committee meeting minutes. Coulthard (2013: 448) identified a number of lexical and collocational features in the disputed email that were also found in other emails and texts written and spoken by Mr. Widdowson, such as *under attack, rhumours* (sic), *disgruntled employees* or *competitors* and *fully expensed.* The evidential strength of these items was reinforced by the finding that these words were not found in the known documents for any of the other candidates (or anyone else in the 190,000 email database). Coulthard (2013: 458) concluded that 'significant lexical choices in the questioned email are consistent with choices Widdowson makes elsewhere' and that 'these coselections do not occur in emails sent by anyone else and so are distinctive'.

In Coulthard's linguistic analysis in the Jenny Nicholl case outlined in the Introduction (Chapter 1), many of the features which he used to identify David Hodgson as the author of the texts (rather than Nicholl herself), were lexical variables, such as *me* and *meself* for *my* and *myself*, the abbreviation *im* for *I am*, and the use of *cya/cu, fone/phone, shit/shite, of/off, iv/ave* and the use of *aint*. A similar text message case is reported by Grant (2013: 467) in which a husband, Christopher

Birks, had murdered his wife, Amanda Birks, and attempted to disguise the timing and mode of her death by sending a series of SMS text messages from her phone. Grant identified consistent stylistic differences between the text messaging style of the two potential authors, many of which were lexical: *ad* for *had*, *bak* for *back*, *wud* for *would*, *dnt* for *don't* and *thanx* for *thanks*. On the basis of these features, and subsequent statistical analysis, Grant (2013: 485–6) arrived at the opinion that the two authors had distinctive styles, that some of the disputed messages were stylistically consistent with Christopher Birks' known messages, and different from Amanda Birks'. Following the presentation of linguistic and non-linguistic evidence in court, Christopher Birks changed his plea from 'not guilty' to 'guilty' and is now serving a life sentence (Grant 2013: 494).

In comparison with the stylometric analyses discussed below, work of this stylistic nature has received less research attention. This may be a result of two factors. First, it might be that although stylistic approaches are widely applied in forensic case work, the reports of such cases are not subsequently published, either because the linguist in question chooses not to, or because legal reasons of confidentiality make it difficult to do so. Second, the pool of practitioners with the skill sets to undertake this kind of stylistic analysis is far smaller than that of computational scientists or computational linguists with the expertise of designing experiments and testing automated algorithms, using a wide range of non-forensic texts, such as literary works, blogs, and news media texts.

2.2.2 Lexis in stylometric approaches to authorship attribution

In stylometric authorship studies, the relative frequencies of a wide range of linguistic features have been used to quantify style, such as syntactic parts-of-speech categories (e.g. Baayen et al. 1996; Hirst and Feiguina 2007) and structural elements including greeting/farewell text, paragraph length, emoticons, hyperlinks and HTML tags (e.g. de Vel et al. 2001; Abassi and Chen 2005; Rico-Suayles 2011). However, these studies aside, lexical variation and lexical features have dominated stylometric authorship research. In particular, the relative frequency of function or grammatical words has received a great deal of research attention, as these words are thought to be 'context-free' (Holmes 1994: 90) and 'topic independent' (Koppel et al. 2009: 11). Although not the earliest (e.g. Ellard 1962) Mosteller and Wallace's (1964) investigation into the authorship of *The Federalist Papers* is widely acknowledged as one of the seminal papers in authorship attribution. In this study, the frequencies

of function words including prepositions, conjunctions and articles, were used as discriminators between the writing styles of the three potential authors: Alexander Hamilton, John Jay and James Madison. This paper was also pioneering in its use of powerful statistical measures, as numerical probabilities were used to express the 'degrees of belief about propositions such as 'Hamilton wrote paper No. 52' (Holmes 1994: 90). On the basis of these function word results, Mosteller and Wallace (1964: 306) were able to make strongly evidenced conclusions about the authorship of disputed political essays, many of which have been supported and confirmed by more recent research (Kjell 1994; Holmes and Forsyth 1995; Tweedie et al. 1996).

Fifty years later, function words are still being used as a marker of authorship in combination with statistical techniques. However, now, they often operate incognito under the guise of 'most frequent/common words', which inevitably comprise function words in any dataset. The work of John Burrows uses such an approach in the analysis of literary texts. Burrows (2002; 2003; 2005) sets out to attribute literary works to their correct authors using lists of the most common words in the datasets. He uses the 'Delta' statistical procedure in categorising authors, which measures how far individual authors or texts diverge from the mean relative frequency of these most common words. He concludes that 'with texts of 1,500 words or more, the Delta procedure is effective enough to serve as a direct guide to likely authorship', especially when using 120–150 of the most common words as markers of style (Burrows 2002: 276).

Besides Delta, frequent words have been used as linguistic features with a wide range of statistical techniques. Grieve (2007) for example, uses the chi-squared statistic to evaluate a series of linguistic features in their ability to attribute *Telegraph* newspaper columns to their author. Chi-squared compares the frequencies of linguistic features in a disputed text with the frequencies that would be expected if the text were written by a particular author, based on the evidence in their known texts (Grieve 2007: 255). Grieve found that accuracy rates of up to 90% were returned when high frequency words were used to discriminate between two possible authors.

Another statistical technique widely used is discriminant function analysis, which involves predicting which author a text belongs to based on a range of 'predictor' linguistic variables. Holmes et al. (2001) use discriminant function

analysis and lists of the 25–60 most frequently-occurring words to determine the authorship of the *Pickett Letters* published during the American Civil War, with authorship disputed between General George Pickett and his widow. They conclude that the published letters were composed by Pickett's wife, countering the claims of historians who believe them to be genuine. Similarly, Can and Patton (2004) use the most frequently occurring words in their study of newspaper and novel writing of two Turkish writers over time. They report that using discriminant function analysis, they achieve an average of 92% accuracy in correctly categorising texts as being either 'new' or 'old' writings of these authors.

Modern stylometric research has seen an increase in the use of machine learning techniques to classify texts and identify authors on the basis of common words. In machine learning methods, the known writings of authors are considered as a set of 'training' documents which are used to train a classifying algorithm. This algorithm identifies linguistic features (or 'vectors'), the relative frequencies of which discriminate between texts or authors, and is then used to assign anonymous or unseen documents to their correct authors based on these features (Koppel et al. 2013: 319). Argamon and Levitan (2005) compare the effectiveness of frequent function words against frequent word pairs. Employing machine learning techniques to discriminate between pairs of authors, their experiments performed on twenty novels found that the 200 and 500 most frequent words 'gave results that clearly show a superiority of function words over collocations as stylistic features' in text classification. Zheng et al. (2003; 2006) focus on authorship in online contexts, in English and Chinese email messages and web forum posts. Their studies make use of 122 function words and a range of machine learning techniques to achieve average authorship prediction accuracies ranging between 70% and 97% (Zheng et al. 2003: 71; Zheng 2006: 378). Kucukyilmaz et al. (2008) study Turkish, and include 78 Turkish function words in their repertoire of stylistic features for analysis. In a corpus of over 200,000 online chat messages, using a number of different machine learning algorithms, they report that among 100 authors, the identity of an author is correctly predicted with 99.7% accuracy.

The breakthrough into using content words as well as function words in stylometric authorship analysis can be credited to David Hoover. Hoover's research (Hoover 2001; 2002; 2003; 2004) utilises cluster analysis and 'most frequent words' in classifying texts and authors. Cluster analysis is a statistical procedure which

'clusters' texts together which are similar to each other in terms of the linguistic features used. In his analysis of literary prose, Hoover (2001) extends the 'most frequent words' used in cluster analyses from the 150 Burrows (2002; 2003; 2005) used, to as many as 500. He claims that after the 200th most frequent word, the majority of words are content words, and argues that 'adding some extremely frequent words that are not function words seems reasonable, under the assumption that their use may also be unconscious' (Hoover 2001: 424). Across all of the experiments in his study the 500 most frequent words of all kinds outperformed purely function words. Hoover (2004: 470) tests the effectiveness of Burrows' Delta on prose corpora and confirms Burrows' (2003) assessment that the accuracy of tests increases as the number of frequent words included in the analysis increases. Indeed, Hoover (2004: 470) notes that the best results in his prose experiments were achieved when using the 600–700 most common words in the corpus:

> Although the traditional view has been that only the most frequent words, typically function words, are likely to be beyond the author's control and are therefore suitable for authorship attribution, the analyses above show that Delta's accuracy continues to increase, at least up to the 600 or 700 most frequent words, at which point almost all the words are content words.

Following Hoover, Diederich et al. (2003) use all of the words in their German newspaper texts as input features in machine learning algorithms. They report accuracy rates of up to 80% in identifying the correct author of a text, and conclude that, 'there is no need to select specific features, we may simply take all word frequencies' (Diederich et al. 2003: 15). In another study, Jockers and Witten (2010) run machine learning experiments in identifying authorship of the *Federalist Papers* using the full lexical repertoires in the texts, as well as two-word strings (word bigrams). For their most successful machine learning technique, which returned a 100% success rate for classification, they present the 50 'most important' features in the classification, which were a combination of function words (e.g. *a, and, the),* content words *(america, confederacies, independent*), and function word + content word bigrams (*arms and, of america, treaties and*). Similarly, Labbé (2007) measures the 'intertextual distance' between texts by considering all of the tokens found within them. Using this method he was able to differentiate between authors of over fifty 10,000 word extracts of English. Moreover, he found that it was the

rare, low frequency, words in the corpora which created the greatest distance between texts written by different authors (Labbé 2007: 48).

Some of the most recent authorship work takes into account all of the data in the text but does so in terms of character n-grams, strings of n characters, so that 'each text is considered as a mere sequence of characters' (Stamatatos 2013: 427). For example, the word *finally* can be represented in character three-grams as *fin, ina, nal, all, lly,* and *ly_* (space). Proponents of this approach argue that character n-grams hold several advantages, such as that they are easy to measure and extract, they are language independent, and that they capture aspects of both content and style (Koppel et al. 2011: 86; Stamatatos 2013: 428). Stamatatos (2013: 343) in his attribution of articles written and published in *The Guardian* newspaper, found that 'character *n*-grams produce models more effective and robust than those based on word features', achieving 90% accuracy. Similarly, Luyckx and Daelemans (2011: 43) attribute samples from three different datasets (two Dutch and one English), finding that 'character trigrams outperform the other feature types in the three data sets' with an accuracy of approximately 80%. Finally, Koppel et al. (2011: 93; 2013: 325) use character four-grams to attribute 500 word samples of blog posts, finding that the method can attribute snippets of this size to one of 1,000 authors with a precision of 93.2%. These character-level n-grams are essentially splitting words up, and so capture the same stylistic information as whole words do. They are attractive to computationalists because splitting words up in this way produces a much higher number of features for comparison (as shown by the *finally* example). The apparent success of these character n-grams, however, is offset by the fact that they are linguistically meaningless, insofar as that the statistical results they produce are more difficult to explain in linguistic terms than is the case for whole words or word clusters (see Section 2.2.3).

What this range of literature shows is the preoccupation that both stylistic and stylometric authorship attribution research have had with lexical variables. Whether in the form of character strings, function words, content words or entire vocabularies, authorship analysts have relied primarily on authors' lexical choices to distinguish between individuals and attribute authorship of disputed documents. Whether the approach is qualitative or quantitative in its emphasis, and regardless of statistical technique employed, the identification of individuals has been done on the basis of lexical variation. Despite this shared focus, stylistic and stylometric

approaches to authorship remain, for the most part, completely independent of each other. The reality is that both approaches have their respective benefits and drawbacks, both methodologically and theoretically. The next section discusses these benefits and drawbacks and highlights how the advantages of one may help address the disadvantages of the other. In turn, an alternative method of analysing lexical variation, in particular the sequential co-selection of lexical items, is suggested as a means by which aspects of both approaches can be combined.

2.2.3 Benefits and drawbacks of stylistic and stylometric approaches

The advantages and disadvantages of stylistic and stylometric approaches to authorship attribution relate to three main points, (i) objectivity and reliability, (ii) theoretical and linguistic validity and explanation, and (iii) accessibility to lay judges and juries.

Qualitative stylistic approaches are often criticized for being too subjective, and thus unreliable, in providing evidence in forensic contexts. Much of this criticism comes from the United States, where the admissibility of expert evidence is determined in relation to the standards of the *Daubert Criteria* which were established at the end of the case of Daubert vs. Merrell Dow Pharmaceuticals, Inc (1993). The criteria were set up to ensure that expert evidence offered is 'scientifically valid', that is:

1. Whether the theory offered has been tested;
2. Whether it has been subjected to peer review and publication
3. The known rate of error; and
4. Whether the theory is generally accepted in the scientific community

(Tiersma and Solan 2002: 225)

McMenamin (2002: 166), as a major proponent of the stylistic method, argues that such an approach satisfies the Daubert criteria. However, many commentators on stylistic approaches argue that they fail to meet the standards of scientific admissibility. Chaski (2001; 2005) is one of the main critics of the stylistic approach. She (Chaski 2005: 2) states that without databases to ground the significance of linguistic features or style markers identified in a stylistic analysis 'the examiner's intuition about the significance of a stylistic feature can lead to methodological subjectivity and bias' (Chaski 2005: 2). This view is shared by

Howald (2008: 235–6) who asserts that from a theoretical point of view, stylistics is 'set up to be weaker in terms of addressing the question of idiolect' and that 'the fact that some practitioners do not validate their results is itself telling of methodological weakness'. Solan (2013: 557), in his measured discussion of stylometric and stylistic approaches, highlights that nobody can be sure that a stylometric approach to a case will fare better than a stylistic one. He adds, however, that practitioners in the stylistic camp 'conduct little or no laboratory work' and 'the result is a dearth of serious research, provoking reasonable questions about the legitimacy of the conclusions reached'. Grant (2013: 491) summarises the major perceived methodological weakness of stylistic analysis as being an 'overreliance on subjective expertise', and Nini and Grant (2013: 176) argue that when stylistic contrasts are drawn in such ways, they are 'loosely defined and can be harder to measure and evaluate'.

In contrast, stylometric approaches are considered to be more objective, empirical, replicable, and ultimately more reliable than their stylistic counterparts. As Solan (2013: 574) notes, computational and stylometric scientists 'are accustomed to testing their algorithms to see how well they work and reporting the rate of error'. Nini and Grant (2013: 176) point out that one of the main disadvantages of stylistic approaches is that they begin with the analysis of texts and the search for distinctive linguistic features, which will be different for every case. In turn, this makes it hard 'to replicate the analysis and therefore to claim objectivity and universality'. In contrast, stylometric approaches either define the markers that they are drawing on (mainly lexical measures, as discussed above) *before* the texts are analysed, or the statistical models used identify important features themselves (e.g. Grant and Baker 2001: 76). Such an approach avoids the subjectivities of the analyst and increases the replicability of the methods and the generalisability of the results. This advantage of stylometric approaches satisfies many of Butters' (2012: 354) methodological concerns regarding authorship attribution—such as the strength, validity and reliability of linguistic features used—in his discussion of ethics and best practices in the provision of forensic linguistic expert testimony.

While quantitative stylometric approaches to authorship analysis appear to offer more objectivity and reliability than the stylistic approach, the main advantage of the latter is its foundation in the linguistic theory of language variation and idiolect. When discussing the statistical focus of stylometric studies, Grant (2008:

226) argues that in forensic contexts 'there are obvious dangers in computationally pursuing an algorithm which distinguishes authors and yet has no linguistic explanation or validity'. He continues:

> In the computational discipline of text mining it might be reasonable to sacrifice linguistic validity in the rush to discovery of an authorship algorithm, but in the forensic field the analyst must be able to say why the features they describe might distinguish between two authors in general, and why they distinguish between the particular authors of the case.

Stylometric studies of authorship rarely, if ever, are able to provide such linguistic explanations as to why authors in their study vary in their use of the particular linguistic features being examined. For example, they may find that the relative frequency of a particular set of function words is able to correctly assign a disputed text to its author, but they are not able to explain why authors vary in the use of these words. Howald (2008: 235) is a proponent of stylometric approaches, but comments that features such as function words 'are not well understood in terms of their link to notions of idiolect'. This is a concern now being voiced from computationalists themselves. Argamon and Koppel (2013: 299) argue that the linguistic features used in authorship analysis 'should enable clear explanation of the stylistic difference' between authors. However, they comment that:

> developments in machine learning and computational linguistics over the last fifteen to twenty years have enabled larger numbers of features to be generated for stylistic analysis. However, in almost no case is there strong theoretical motivation behind the input feature sets, such that the features have clear interpretations in stylistic terms
>
> (Argamon and Koppel 2013: 300)

Similarly, Stamatatos (2013: 428), himself an advocate of the character n-gram approach, comments on precisely this issue:

> they [character n-grams] capture small pieces of stylistic information, making the interpretation of the stylistic property of text very difficult if not impossible. Such an interpretation is crucial in case the authorship attribution technology is used as evidence in a judicial process.

It is here where stylistic analyses hold an advantage. Stylistic analysis traditionally puts qualitative analysis at the forefront, and the qualitative study of writing does not just focus on *what* forms are used by a writer, but also *how* and *why* they are used (Johnstone 2000: 35). Similarly, forensic stylistics gives qualitative evidence

primacy over quantification. As McMenamin (2010: 491) argues, 'linguistic assessments of style precede their expression as numerical values and are often a more realistic representation of the facts'. Forensic stylistics has its foundations in language variation, and involves the analyst identifying 'style markers' in a text, which McMenamin (2010: 488) defines as an author's 'choice from optional forms', which are 'the observable results of the habitual and usually unconscious choices an author makes in the process of writing'. He (McMenamin 2002: 47, 53) argues that such linguistic choices authors make reflect their individual linguistic competence and their unique combination of linguistic knowledge, cognitive associations and extra-linguistic influences: their idiolect. As Nini and Grant (2013: 176) add:

> stylistic methods do provide a justification on why their markers distinguish authors. The differences in spelling, word forms, grammatical constructions and so on originate from the different sociolinguistic background that each individual presents.

In addition to these methodological and theoretical discussions with regard to stylometric and stylistic approaches, there is another important factor which is far less regularly commented upon. There is a vexed relationship between the method used in a forensic case, and the accessibility of the results obtained to the lay decision-makers in legal contexts, judges and juries. Clark (2011: 11) questions whether stylistic analyses 'are sufficiently reliable and specialized that they actually help the jury decide anything that they could not decide for themselves'. On the other hand, McMenamin (2002: 129) argues that 'in the courtroom, qualitative evidence is more demonstrable than quantitative evidence because it is the language data that are presented'. In turn, he claims, 'qualitative results appeal to the nonmathematical but structured sense of probability held by judges and juries'. This is a view shared by Cheng (2013: 547) who comments that jurors are more comfortable weighing up qualitative evidence, and that statistical models rely on mathematical assumptions with which it is unrealistic to expect jurors to engage. At the same time, Argamon and Koppel (2013: 300, 315) state that without a firm basis in linguistic theory, it is not possible to establish or explain algorithmic evidence for a proposed attribution, which in turn makes the task of conveying the importance of computational results to non-experts very difficult.

Therefore, overall, on the one hand stylometric approaches boast higher levels of reliability, objectivity and replicability, and on the other stylistic

approaches produce results that are linguistically explainable and underpinned by linguistic theory. Recent years have seen a small number of studies set out to combine the best aspects of these two approaches.

2.2.4 Combining the stylistic and the stylometric

The emerging trend in authorship analysis research is to move towards a situation in which stylometric and stylistic approaches are combined, producing reliable quantitative results with clear linguistic underpinning. The main way in which this has been pursued is by selecting features for analysis which capture aspects of an individual's idiolect. Chaski (2001; 2005) makes use of a set of features she describes as 'syntactically classified punctuation'. In this approach, punctuation marks (commas, colons, exclamation points etc.) 'are counted by the kind of boundary or edge which the punctuation is marking' (Chaski 2005: 6). Chaski (2005: 12) reports an overall success rate of 95.7% when attributing texts between 200 and 600 words in length using discriminant function analysis. Similar results have been reported elsewhere (Chaski 1997; 2001; 2004). Chaski (2005: 3) claims that the primary difference between her method and other computational stylometric methods is 'the syntactic method's linguistic sophistication and foundation in linguistic theory'. As Nini and Grant (2013: 176) point out, Chaski 'does not state, however, why there should be difference in syntactic behaviour between individuals'. Furthermore, throughout Chaski's reports, the reader is not offered any examples of syntactically classified punctuation which are particularly distinctive or idiolectal of individual authors in her studies.

Turell (2010) and Queralt and Turell (2012) also focus on syntactic markers and also use discriminant function analysis to classify disputed texts with their correct authors. The syntactic elements they focus on are part-of-speech bigrams (combinations of two categories) and trigrams (combinations of three categories) such as PREPOSITION + DETERMINER or PREPOSITION + DETERMINER + NOUN. Turell (2010: 235–237) reports an accuracy rate of 63% for syntactic trigrams and 83.4% for bigrams. Based on these results, she concludes that 'sequences of grammatical categories observed in a writer's "idiolectal style" can be used quite reliably as valid markers of authorship'. She adds that these markers 'seem to reflect the writer's preference for a certain formulaic textual style and for specific combinations of linguistic categories' (Turell 2010: 238–9). Again, however, Turell's work does not

provide any examples of which types of syntactic n-grams are distinctive of individual authors.

Grant (2010; 2013) combines the stylistic and statistical by first identifying potentially distinctive style markers through a quantitative analysis of SMS text messages. He then demonstrates how the level of similarity between known and questioned documents can be measured on the basis of these features using Jaccard's similarity coefficient (a statistic used in this thesis, see Section 4.4.3). Grant (2013: 472) claims that the aim of his approach is to 'demonstrate how it is possible to derive a methodologically rigorous approach to stylistic authorship analysis that can result in statistically described results'. By combining the quantitative identification of style markers with statistical testing, Grant (2013: 484) is able to claim that the disputed texts in question show a statistically significant difference from the known texts of one candidate author (who did not in fact write them) but no equivalent difference to the known texts of the other (who did write them).

Finally, Argamon and Koppel (2010; 2013) and Nini and Grant (2013) have both drawn on the theory of Systemic Functional Linguistics (SFL) (Halliday and Matthiessen 2004). SFL is a useful theory in relation to authors' stylistic choices in that within it, 'grammar is a network of possible choices, with more general or abstract choices constraining which more specific choices are allowed' (Argamon and Koppel 2013: 302). Within an SFL framework, the stylistic choices authors make from within the overall system at their disposal are determined by their social backgrounds and their interpretation of the context in which they find themselves. In turn, these choices reflect their idiolectal preferences, or (in SFL terms), their 'codal variation' (Argamon and Koppel 2013: 303; Nini and Grant 2013: 179). While Argamon and Koppel (2010; 2013) use SFL in conjunction with machine learning techniques to address the problem of 'author profiling' (see Section 2.4), Nini and Grant (2013) apply this approach in authorship attribution. They code the academic essay writing of three undergraduate students for variables relating to clause complexity, conjunctions, modality, mood, nominalisation, theme and transitivity. They use ANOVA (analysis of variance) to compare the mean frequencies of these variables across the three authors to identify any significant differences in frequency of use. They found that the three students significantly varied in their use of determiners *vs* pronouns, elision of nominal groups and relational transitivity patterns. In addition to these statistical findings, Nini and Grant (2013: 187) present

a number of textual examples to qualitatively demonstrate this variation in use across their three authors. They argue that the variation itself is:

> generated by the different experiences that the individuals have of that genre, in this case, of the different readings and academic writing that they had done before, the different approaches to academic writing that they have been taught, etc.

(Nini and Grant 2013: 188)

These studies which aim to develop methods which combine the best aspects of stylometric and stylistic methodologies represent promising advancements in the field. They are a rejection of the view that these different standpoints are incompatible (Howald 2008: 235) and focus on bridging the gap between them. The combined advantages aim to produce methods in which (i) there is a clear theoretical motivation for the linguistic features being drawn on in the comparison of authors, (ii) that the similarities and differences between authors, and any subsequent attribution of disputed texts, are based on reliable and replicable statistical techniques, and (iii) that the statistical results produced can be explained and described in linguistic terms. Essentially, such an approach provides the analyst with both quantitative and qualitative evidence for identifying idiolectal differences between authors and identifying the author(s) of disputed texts. This thesis continues this trend of combining the stylometric and the stylistic, and uses corpus linguistics as the methodology with which to achieve this combination. Corpus linguistics has been applied in various fields of linguistic enquiry (from Critical Discourse Analysis to language teaching) in such a way that quantitative and qualitative results can be used to complement one another (McEnery and Wilson 2001: 76–77). In applying a corpus methodology, this thesis puts at the centre of its analyses lexical variation, an aspect of linguistic variation that has been at the forefront of stylistic and stylometric studies alike. In particular, the focus here is on authors' co-selection of lexical items across texts, and how co-selection preferences and collocation patterns reveal idiolectal variation.

2.3 Collocation, idiolect and word n-grams

The use of lexical variation in stylometric analyses in authorship attribution research, as outlined above, has most often been limited to focusing on the relative frequencies of words across authors and texts. While most of these studies have reported good results, their results are not explainable in stylistic terms, or in terms of a linguistic theory of idiolect. This thesis aims to remedy that, and proposes a use of lexical variation that is underpinned by idiolectal theory. In particular, focus here is on sequences of lexical items, or 'collocations', 'the characteristic co-occurrence patterns of words' (McEnery and Wilson 2001: 85).

Coulthard (2013: 447–8), in discussing the difficulties of dealing with topic-sensitive content words in authorship attribution, emphasises the usefulness of collocations:

> Given the same basic topic, different speakers/writers will still choose to mention and/or omit different aspects and choose differing lexis to encode any given topic item. Thus, while the occurrence of individual lexical items shared between topically related texts is significant in authorship attribution, much more significant is the shared occurrence of coselected items or what linguists call *collocates* [...] Thus, we can see clearly that, although in theory anyone can use any word at any time, the topics they choose, the aspects of the topic they decide to focus on, and their preferred linguistic realizations ensure that texts quickly become linguistically unique.

This is not the first time Coulthard has commented on the distinctiveness of sequential lexical collocates. Coulthard (2004: 441) empirically demonstrates the evidential value of lexical strings. Using Google and the web as a reference corpus, he is able to show that as the length of a lexical string increases so too does its rarity, to the extent that it fairly quickly becomes entirely unique to the speaker/writer (Figure 1). He found that lexical strings of six or seven words in length produce either zero, or very few, results from the billions of documents Google searches within. He argues that 'rarity scores like these begin to look like the probability scores DNA experts proudly present in court' (Coulthard 2004: 442). Similarly, lexical strings have been used as signifying individual language variation in plagiarism research, a sister field of authorship attribution. Writing on idiolect in relation to plagiarism, Johnson and Woolls (2009: 112) outline two assumptions that deal with word sequences and selection. First, 'most sequences of words are unlikely to be selected and arranged in the same order by two individuals, whether writing on

Figure 1. The distinctiveness of lexical sequences (from Coulthard 2004: 441)

String	Instances
I picked	1,060,000
I picked something	780
I picked something up	362
I picked something up like	1
I picked something up like an	0
an ornament	73,700
like an ornament	896
something like an ornament	2
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
if I could	2,370,000
if I could carry	1,600

the same topic or not'. Second, 'extended common sequences are even more indicative of a common source' and therefore their appearance in two texts would suggest that these texts share the same author.

Also within a plagiarism context, a similar argument is made by Culwin and Child (2010: 16) in their investigation of possible student plagiarism, who claim that:

> the occurrence of a string of as little as six words in a student submission, whose frequency of occurrence is shown by an Internet search to be unique or nearly unique, can be assumed to be, beyond reasonable doubt, copied.

The analysis of collocation patterns in texts has been a major focus of corpus linguistics for the last half a century (McEnery and Hardie 2012: 122). Collocation as a concept was first introduced by Firth (1957: 179) who famously wrote 'you shall know a word by the company it keeps', and was later developed by a number of key Neo-Firthian scholars (Sinclair 1991; Louw 1993; Stubbs 2001; Hoey 2005) and continues to be a main route of investigation in the field (e.g. Gries 2013). It is now commonly believed that the associations that people build between words, and how they reproduce these associations in collocations is a psychological phenomenon (Mollin 2009a; Michelbacher et al. 2011; Gries 2013). Many theories and frameworks have developed from corpus linguistics and psycholinguistics which aim to describe how individuals store and produce collocations and word

sequences. For example, Nattinger and DeCarrico (1992: 1) coin the term ‘lexical phrases’ as being ‘chunks of language of varying length’ existing ‘somewhere between the traditional poles of lexicon and syntax’ and which ‘occur more frequently and have more idiomatically determined meaning than language that is put together each time’. This is related to Sinclair’s (1991: 109) ‘idiom principle’ which holds that a language user ‘has available to him or her a large number of semi-preconstructed phrases that constitute single choices’. This, Sinclair argues, works simultaneously with the ‘open-choice’ principle, in which an utterance is generated on a word-by-word basis. A main proponent of the idea that lexical items are stored in clusters and reproduced as such is Wray (2002; 2009) in her theory of formulaic language. She uses the term ‘formulaic sequence’ to describe:

> a sequence, continuous or discontinuous, of words or other elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar
>
> (Wray 2002: 9)

Alongside these theories that collocations and phrases are stored in clusters in the mind and reproduced as such by writers is Hoey’s (2005; 2006) theory of ‘lexical priming’, also discussed and developed at length by Pace-Sigge (2013). Alongside these theories that collocations and phrases are stored in clusters in the mind and reproduced as such by writers is Hoey’s (2005; 2006) theory of ‘lexical priming’, also discussed and developed at length by Pace-Sigge (2013). Hoey (2005: 5) argues that collocations are ‘a psychological association between words’ and that ‘we can only account for collocation if we assume that every word is mentally primed for collocational use’ (Hoey 2005: 8). He claims that as language users encounter a word in speech and writing, it ‘becomes cumulatively loaded with the contexts and co-texts in which it is encountered’. The result of this, he continues, is that our knowledge of that word ‘includes the fact that it co-occurs with certain other words in certain kinds of context’ (Hoey 2005: 8). Overall, he (Hoey 2006: 53) asserts that:

> If we argue that language is a choice (as we must), then what is chosen is not just the lexical item and the grammar, but the lexical item and its collocations, colligations and semantic associations.

The theory of lexical priming is extended by Hoey beyond individual words being primed with others. He holds that the same applies to word sequences which are built from these constituent items, so that word sequences too become loaded with contexts and co-texts with which they occur, a process he calls 'nesting' (Hoey 2005: 8). In the same way as Sinclair's (1991) idiom principle, Nattinger and DeCarrico's (1992) lexical phrases and Wray's (2002) formulaic sequences, such 'nesting' of word sequences in the minds of language users, and the production of these sequences, serve to 'simplify the memory's task' (Hoey 2005: 8).

These concepts of lexical priming and nesting are closely related theoretically to usage-based theories of linguistic output, and in particular 'Exemplar theory'. Exemplar theory was first applied in linguistic research in relation to analysis of phonetic variation (e.g. Pierrehumbert 2001 and more recently Hay and Bresnan 2006; Foulkes and Hay 2013) and has also been applied to the cognitive organization and representation of syntactic construction (e.g. Bybee 2006; Walsh et al. 2010). In exemplar models, every 'token', or instance of linguistic experience, is stored as a separate 'exemplar', is classified as belonging to a particular 'category', and is placed in a vast organizational network where each category is represented in memory by a large cloud of remembered tokens of that category (Pierrehumbert 2001: 140; Bybee 2006: 716). As Pierrehumbert (2001: 140) explains:

> These memories are organized in a cognitive map, so that memories of highly similar instances are close to each other and memories of dissimilar instances are far apart. [...] For example, a recollection of the phrase *Supper's ready!* could be labelled as "Mom" and "female speech", in addition to exemplifying the words and phonemes in the phrase. [...] When a new token is encountered, it is classified in exemplar theory according to its similarity to the exemplars already stored.

Such an exemplar-based approach deals not only with the storage of linguistic information, experiences and memories, but also with the production of language. When producing language, individuals delve into their cloud of stored linguistic information and activate the relevant exemplars. Specifically with regard to collocation, and in a similar way to lexical priming, when one word is selected by an author, its subsequent collocates will be selected based on their previous experience of this word. Recently, and most relevant here, is Barlow's (2013) extension of Exemplar theory into discussions about lexical strings. He (Barlow 2013: 473) argues that Exemplar theory, and other usage-based models of language, offer 'the

most promising framework' for explaining the relationship between language use and cognition, and thus the subsequent production of lexical strings and collocations.

A point should be made regarding the implications that non-native language learning has for these theories of word association and collocation production. Wray (2002) makes a number of observations about the role of formulaic sequences in the acquisition of a second language. She argues that formulaic sequences are 'a great support in the early stages of L2 acquisition for most kinds of learner' (Wray 2002: 147). More specifically, she claims that even very young L2 learners are adept are employing formulaic chunks of language to achieve particularly manipulative speech acts and to access the inner workings of the language in question. Similarly, she argues that the desire to communicate prompts adult L2 learners to adopt formulaic sequences, and 'fossilize' any useful lexical string in the L2 language once it has been found to work in achieving communicative goals (Wray 2002: 148). The situation is seemingly more complex for lexical priming, however. Hoey (2005: 183) highlights that when the lexis of the first language is being primed it is happening for the first time, but when a second language is learnt, 'the primings are necessarily superimposed on the primings of the first language'. Therefore, the learner will automatically activate the primings from their first language, and so their semantic associations of this new word 'will be the same, or at least very similar to, those of the L1 equivalent' (Hoey 2005: 183). The ultimate challenge this poses to learners, Hoey (2005: 184) argues, is that it will result in words that mean exactly the same in the two languages being primed differently for L1 learners as regards collocations, colligations and semantic associations (Hoey 2005: 184). This challenge also arises in discussions of second language acquisition and Exemplar theory. Hall and Boomershine (2006: 5) note that second language acquisition becomes increasingly difficult with age. They argue that in an Exemplar model, this is because tokens, categories and robust exemplars of individuals' first language are already cognitively stored, much like the initial L1 primings Hoey (2005: 183) refers to. In their experiments in which they examine the ability of adult native speakers of a language to perceive non-native phonetic stimuli, Hall and Boomershine (2006: 25) confirm this hypothesis, finding that when exposed to Greek phones, native English speakers make use of their English exemplars in interpreting them. However, the experiments also find that over time native English listeners are able

to recognise difference between Spanish phones which are not contrastive in English, and recognise similarities in pairs of Spanish phones that are natively contrastive in English. They argue that this is evidence to suggest that given enough experience and encounters with non-native language stimuli, 'even speakers who begin learning the language after their first language categories are in place can change their grammars to reflect the second language categories' (Hall and Boomershine 2006: 26). On the basis of these results, they (Hall and Boomershine 2006: 26) conclude that under an Exemplar-based model second language acquisition can progress toward native-like fluency with enough exposure in the second language and, importantly, 'the "critical period" for second language acquisition must indeed be considered to be "life"'. Despite dealing with phonology rather than collocation, the conclusions of these experiments might reasonably be extended to argue that the same processes (will) take place in the acquisition of lexis, word associations and collocations. This would present a more optimistic outlook than Hoey (2005: 183).

What all of these theories have in common—from the idiom principle and formulaic sequences to lexical priming and Exemplar theory—is a belief that, regardless of (non-)nativeness, the storage and production of lexical items and their collocates are *unique* to individual language users. When discussing lexical phrases, Nattinger and DeCarrico (1992: 39–40) state that some may be 'idiosyncratic phrases that an individual has found to be an efficient and pleasing way of getting an idea across'. Similarly, Wray (2002: 101) argues that 'there is not simply a single stock of formulaic sequences which all speakers learn first and then draw upon', but that 'each person, in each unique situation, will apply slightly different selection criteria to a slightly different set of options, from those available to anyone else'. Schmitt et al. (2004: 138) echo this sentiment with regard to the individuality of formulaic sequences. They argue that it is idiosyncratic to individual speakers whether they store particular clusters or not. They assert that as part of a person's idiolect they will 'have their own unique store of formulaic sequences based on their own experience and language exposure'. When discussing lexical priming, Hoey (2005: 8–15) draws on Firth's (1957) notion of 'personal collocations', emphasising that 'an inherent quality of lexical priming is that it is personal' and that 'words are never primed *per se*; they are only primed for someone'. He goes on to explain that:

> Everybody's language is unique, because all our lexical items are inevitably primed differently as a result of different encounters, spoken and written. We have different parents and different friends, live in different places, read different books, get into different arguments and have different colleagues.
>
> (Hoey 2005: 181)

He (Hoey 2005: 9) aligns his theory of lexical priming with Hopper's (1988; 1998) theory of 'emergent grammar', which holds that grammar is the output of 'routines', the repeated use of collocational groupings which result in the creation of a distinct grammar for each individual. Again, such grammars are different from person to person because 'every speaker's experience and knowledge of routines is different' (Hoey 2005: 9). In the same way, because the usage-based Exemplar theory hinges on language users' stored memories of linguistic experiences, which will differ from person to person, the individual's linguistic output based on these exemplars will also be unique. Indeed, this is one of the explanations Barlow (2013: 473) offers when discussing why individuals' language use can be distinct from others within their discourse communities. Finally, Coulthard (2004: 440) discusses Sinclair's (1991) open-choice and idiom principle, arguing that because both principles work side-by-side, any short string of words may be produced either as 'an idiosyncratic combination or a frequently occurring fixed phrase'.

These theories and arguments emerging from corpus linguistics and psycholinguistics all suggest that collocations and collocational patterns observable in a person's writing are manifestations of that person's distinctive idiolect. Indeed, there is existing research from corpus linguistics that has used collocations as a means of identifying and analysing individual's idiolects. Mollin (2009b), for example, employs a robust corpus-based statistical approach to analysing idiolectal collocations of the former Prime Minister of the UK, Tony Blair. Through a comparison of a corpus of Tony Blair's speech/text with the British National Corpus she claims to identify sixteen 'maximiser collocations' (Mollin 2009b: 367) (e.g. *absolutely committed, entirely accepted*) as being idiosyncratic preferences, or 'Blairisms'. In another corpus study, Barlow (2010: 3) focuses on the speech of five White House press secretaries in transcripts of press conferences across a two year span. He (Barlow 2010: 21) finds that in terms of two- and three-word combinations, for example with *we are, the president* and *I don't know*, intra-author stability contrasts with inter-speaker variability, concluding that 'there are dramatic

differences in the speech of individual speakers across a wide range of lexico-grammatical patterns'.

Collocations and strings of lexical words are referred to in linguistics by a wide range of different names, such as 'concgrams' 'flexigrams', 'lexical bundles', 'multi-word expressions,' 'prefabricated phrases', 'skipgrams' (Nerlich et al. 2012: 50). In authorship attribution, Juola (2008: 265) refers to them simply as word n-grams—lexical strings of *n* words—and describes them as a means by which to take advantage of both vocabulary and syntactic information in texts and as an effective way of capturing words in context. These word n-grams are not to be confused with character n-grams or syntactic n-grams used in other authorship studies. There is a precedent for using word n-grams as features in stylometric approaches to authorship attribution, with varying degrees of success. Hoover (2002; 2003), for example, compares the effectiveness of individual frequent words with that of frequent word sequences in distinguishing literary texts by different authors and grouping texts by the same author. He reports that when frequent word sequences are used instead of, or in addition to, frequent words in cluster analyses accuracy improves, and that 'analyses based on frequent sequences even provide completely correct results in some cases where analyses based on frequent words fail' (Hoover 2002: 157). Also working with literary texts, in particular 353 poems by five different writers, Coyotl-Morales et al. (2006: 9) find that frequent two- and three-word sequences can attribute text to their correct authors with 83% accuracy. In a forensic context, Juola (2013: 295–7) used three-word sequences to successfully attribute a set of anonymously-written antigovernment articles to the person who claimed authorship of them in a deportation case.

In contrast, Grieve (2007: 263) evaluated the success of two and three-word collocations in the attribution of newspaper columns to their correct author and found that they performed poorly. In fact, the three-word collocations were the least successful of the many features tested in his study, while in comparison, character-level n-grams performed far better. This finding aligns with that of Sanderson and Guenter (2006: 9) who also found that character sequence analyses generally outperform word sequence features. Most recently, Larner (2014) tests the usefulness of formulaic sequences (in Wray's [2002] terms) as a marker of authorship in a forensic context based on the assumption that they are less likely to be under the conscious control of authors, and so harder to disguise. He (Larner

2014: 10) identified 301 different formulaic sequences (e.g. idioms, clichés, similes) exhibited in the 100 personal narratives he had collected from twenty different authors, including phrases such as *in the end, at least, go back* and *in fact*. He found that 'texts produced by the same author are more similar in their use of formulaic sequence types than text by different authors' (Larner 2014: 13). However, neither the amount of formulaic sequences, nor specific types of formulaic sequence, was successful in attributing disputed documents.

Despite the mixed results of collocations and word n-grams in authorship experiments, the idiolectal nature of word associations and collocate production makes them a promising focus for authorship analysis, and they have received far less research attention than the relative frequencies of individual words. Collocations and word n-gram patterns offer an aspect of linguistic and stylistic variation which appears to be clearly related to an individual's idiolect, and so should prove useful in distinguishing between the writings of individuals and attributing disputed texts to their correct authors. Pursuing these hypotheses is the motivation for Chapters 5 and 6 of this thesis. In addition, however, analysing how words are used in context is also a potential route of enquiry for a younger field of authorship analysis: that of author profiling.

2.4 The emergence of author profiling

Whereas authorship *attribution* is concerned with the identification of individual authors, author *profiling* distinguishes between classes of authors (Rangel et al. 2013: 1) and seeks to determine characteristics of a text's author (Argamon and Koppel 2013: 307). Thus, in contrast to authorship attribution where the focus is on individuals' idiolects, author profiling is concerned with the sociolects of groups. As Coulthard et al. (2011: 538) argue:

> The most successful attempts to profile single texts into sociolinguistic categories do so using complex computational and statistical models, and take into account a large number of linguistic variables.

Over approximately the last ten years, authorship profiling research has gained attention, but much less than authorship identification. Most of the early author profiling work extended the tradition of quantitative sociolinguistics (Labov 1966; Trudgill 1972; Milroy and Milroy 1978) in focusing on the social variable of gender. For example, Thomson and Murachver (2001) found that using discriminant

function analysis it was possible to successfully classify experiment participants by gender with 91% accuracy on the basis of linguistic features such as questions, expressions of emotion, apologies, insults, adjectives, adverbs and conjunctions, Argamon et al. (2003) also identified significant differences between male and female writing in the *British National Corpus* in terms of how frequently they use function words and part-of-speech bigrams and trigrams. In turn, Koppel et al. (2002) utilised these differences in developing a text categorisation technique that could correctly infer the gender of the author of an unseen written document with approximately 80% accuracy. Schler et al. (2005) analysed age as well as gender. Their learning algorithm was able to correctly categorise blog texts by gender with an overall accuracy of 80.1%, and by age with an accuracy of 76.2%. Age and gender continue to be the main variables of interest at the PAN authorship and plagiarism lab held annually as part of the CLEF conference. In 2013 and 2014, the authorship profiling task involved researchers determining the age and gender of the author of a document.

Some research has aimed at identifying other author characteristics, besides age and gender. Using written texts from the *International Corpus of Learner English*, Koppel et al. (2005) used various types of linguistic error to successfully identify the native language of Russian, Czech, Bulgarian, French and Spanish authors writing English with over 80% accuracy. Using the same dataset, Wong and Dras (2011) relied on syntactic parse trees and machine learning techniques to classify texts by the native language of their authors, also reporting approximately 80% accuracy. Van Halteren (2008) used word n-grams as input features for various classification methods in the identification of the source language of translated European Parliament speeches with the highest rates of accuracy achieved being between 87.2% and 96.7%. Identifying the personality type of the author of a text has also received research attention. Drawing on the work of language psychology (e.g. McCrae and Costa 1996; Pennebaker et al. 2003; Cohn et al. 2004), Argamon et al. (2005) found that function word use was good for identifying extraversion in writers (preference for the company of others), and appraisal features (expressing an attitude towards something) were useful for identifying neuroticism (a tendency to worry). Similarly, Luyckx and Daelemans (2008) and Noecker et al. (2013) both use syntactic input features and machine learning techniques to predict authors'

personalities in relation to the Myers-Briggs Type Indicator in a corpus of Dutch student essays.

Most of the recent author profiling research uses sets of linguistic features and attempts to classify and predict multiple discrete author traits and characteristics. Estival et al. (2007) use a range of character-based features (e.g. punctuation, word length), lexical features (e.g. function words, parts-of-speech) and structural features (e.g. HTML tags, paragraph breaks) to provide probabilities of email authors' demographic traits (age, gender, geographic origin, level of education and native language) and five psycho-metric traits (agreeableness, conscientiousness, extraversion, neuroticism and openness). They used various different machine learning algorithms and report prediction accuracy rates of between 53.16% (for predicting agreeableness) and 84.22% (native language). Pham et al. (2009) developed a profiling framework to automatically predict the age, gender, geographic origin and occupation of Vietnamese bloggers. They too used character and part-of-speech features, as well as topic recognition and a number of different machine learning classifiers and reported a classification accuracy of 77% across all four traits. Argamon and colleagues (Argamon et al. 2009; Argamon and Koppel 2013) report on their various applications of author profiling across a range of author traits: gender, age, native language and personality type. They use a combination of corpora (blogs and learner essays) and a set of lexical and syntactic features as inputs into a machine learning technique, with classification accuracy rates between 65.7% (personality) and 82.3% (native language).

The value of such research and findings to forensic analysis is clear. As Rangel and Rosso (2013: 1) argue, being able to identify social characteristics of an author based on textual evidence alone would help in the identification of suspects who may have been responsible for authoring a disputed text. The application of author profiling to narrow the pool of potential suspects is also proposed by Argamon et al (2009: 121) who claim that profiling of this kind can 'help police identify characteristics of the perpetrator of a crime when there are too few (or too many) specific suspects to consider'. However, despite the promising results that the research outlined above has returned, such work 'is not certain enough to provide evidence to the courts' (Coulthard et al. 2011: 538). The reason for this, Coulthard et al. (2011: 538) explain, is:

> Normal scientific method moves from observation of a large number of examples to a generalization, drawing a conclusion about that instance. Profiling involves taking a single example, and by matching it to a well-founded generalization, drawing a conclusion about that instance. This reversal of 'normal' scientific method must be done with caution and is prone to error. The single instance could easily be a statistical outlier such that any generalization might be considered not to apply.

In other words, no matter how much we know from previous research about, for example, male and female tendencies in writing, any suspect in a given case may not be characterised by the patterns and preferences linguists think they have established. It is fair to suggest, for example, that a male author in a given case may use a certain linguistic feature (or collection of features) with a relative frequency that would have him classified as a female; based on the generalisations available to us, this profile would be inaccurate. The risk here is that author profiling research continues along this quantitative trajectory and eventually arrives at the situation which has developed in authorship attribution work, in which a huge number of studies rely on stylometrically derived results which have no linguistic explanation. Steps should be taken in this relatively early phase of authorship profiling research to ensure that quantitative and statistical results are supplemented by qualitative and stylistic analyses, in order to identify the precise nature of the linguistic differences between groups of authors. In order to move towards achieving this complementary approach, and underline why it would be valuable, a useful parallel can be drawn between forensic author profiling and criminal profiling in forensic psychology.

2.4.1 Linguistic profiling and criminal profiling

Grant (2008: 224) claims that despite not being permissible as reliable forensic evidence, linguistic profiles of authors may hold investigative value, and argues that 'different sorts of linguistic evidence may play different roles within the investigative and judicial process'. In his chapter, Grant (2008: 223) draws a parallel between sociolinguistic profiling and psycholinguistic profiling. This link between socio- and psycholinguistic profiling offers a precursor for the more general comparison here between forensic linguistics and forensic psychology or 'investigative psychology'. Broadly speaking, investigative psychology, a term coined by Canter (1995), is the application of psychological research and principles to the investigation of criminal behaviour. One of the fundamental questions that investigative psychology aims to address is, based on the evidence collected in the

case, ‘what inferences can be made about the characteristics of the offender that may help identify him or her?’ (Canter and Alison 2000: 3). Related to this is the process of ‘criminal profiling’, which involves ‘identifying personality traits, behavioural tendencies, geographical location, and demographic or biographical descriptors of an offender’ based on characteristics of the crime (Bartol and Bartol 2008: 82). This process goes by various names besides ‘criminal’ profiling, such as ‘offender’, ‘behavioural’ and ‘investigative’ profiling, all of which broadly describe the same thing. In turn, such behavioural clues left behind in criminal evidence has been drawn upon in the linking of crimes committed by the same individual (e.g. Bennell and Woodhams 2012; Bennell et al. 2014).

The similarities in aims and function of criminal profiling and author profiling are clear. However, as Grant (2008: 224) points out, like author profiles, psychological profiles are rarely admitted as evidence in the UK courts (Ormerod 1999; Ormerod and Sturman 2005). The main criticisms targeted at traditional criminal profiling are that it relies on unsubstantiated theories of personality and behaviour. For instance, they hold that human behaviour is consistent across different situations, and that the more similar any two offenders are socially, the higher the resemblance in their behaviour in any given offence (Alison 2005: 3–4). In addition, criminal profiling research is founded on the presuppositions that crime scene clues provide clues to the perpetrator’s personality traits, habits and thought processes, and that key factors of the personality identified are generalisable across other situations, crimes and individuals (Bartol and Bartol 2008: 94). There are also arguments that the results of profiles, and the advice offered in profiling reports, are often difficult to interpret and evaluate and conclusions arrived at can be ambiguous and unverifiable (Alison 2005: 6; Bartol and Bartol 2008: 95). Overall, though, the main issue is that profilers often rely too heavily on their ‘gut feelings’ and personal judgements, rather than on robust, scientific and empirical exploration (Bartol and Bartol 2008; Canter and Youngs 2009: 7). Such is the extent of methodological and theoretical flaws of traditional criminal profiling that some argue that it should be used with extreme caution in criminal investigations, and ‘not at all as evidence in court, until research demonstrates its predictive validity’ (Alison et al. 2002: 116).

In recent years, however, research in investigative psychology and criminal profiling has developed new statistical and algorithm-based techniques for identifying relationships between criminal behaviour and offender characteristics.

One technique that has gained a lot of momentum is multidimensional scaling, which is mathematically the same as cluster analysis, and represents the relationship between variables (behaviours and characteristics) in a very clearly visual way (Canter and Youngs 2009: 101). It has been used to link criminal behaviours with the geographical location of offenders (Paulsen 2006), as well as their interpersonal personalities, occupation, gender, age, income, marital status and whether they have children (Youngs 2004; Häkkänen et al. 2004; Wachi et al. 2007; Zaitsu 2010). This identification of statistical and quantitative relationships between criminal behaviour and offender characteristics is known as a 'nomothetic' approach to profiling. It 'tries to make general predictions' and searches for 'general principles, relationships, and patterns by examining and combining data from many individuals' (Bartol and Bartol 2008: 95). With these developments and increased methodological rigour, criminal profiling can help in the investigative process, developing a 'manageable set of hypotheses for identifying who *may* have been responsible for the crime' and ultimately serving to 'eliminate large segments of the population from further investigation' (Bartol and Bartol 2008: 83).There is an alternative approach to this nomothetic criminal profiling: an 'idiographic' approach. In contrast to the nomothetic approach, the idiographic approach emphasises the intensive study of one individual. As Turvey (2012: 122) describes:

> Idiographic (individual case) study builds knowledge about the characteristics of a particular case. It is necessary when trying to understand the peculiar characteristics, dynamics and relationships between a particular crime scene, victim and offender. Idiographic offender profiles, therefore, are characteristics developed from an examination of a single case, or series of cases linked by a single offender.

Bartol and Bartol (2008: 95) highlight that 'profiling that relies exclusively on the idiographic approach is in far more danger of missing the mark than profiling based on the nomothetic approach'. Despite this, and the fact that most profiling is nomothetic in nature, they continue: 'many profilers and clinicians prefer exclusive use of the idiographic approach'. Turvey (2012: 122) proposes an argument that may go some way in explaining such a preference. He describes a nomothetic profile as 'an average' or 'prediction' and claims that 'it does not describe a real offender walking around and breathing in the real world'. In contrast, he argues, an idiographic profile is 'concrete' as 'it describes an actual offender who exists in the real world'. Overall, it might be that the idiographic approach offers a qualitative

solution to the abstractness of nomothetic procedures. Instead of relying solely on quantitative generalisations that are difficult to reliably apply to any one individual, an idiographic approach analyses, in detail, the various dynamic elements of evidence available in any given case.

All research in author profiling (discussed above) is nomothetic in approach. In linguistics there are very few case studies of individual linguistic profiles. Mollin (2009b) is an example from corpus linguistics, and some sociolinguistic studies have focused on the language of one individual, usually either a prominent cultural or historical figure. Kredens (2002), for example, compared case studies of two individuals' idiolects: Robert Smith from British rock band *The Cure* and Steven Patrick Morrisey of *The Smiths*. Johnstone (1996; 2009) analysed the linguistic practices of Barbara Jordan, former Texas State senator and U.S. congresswoman. Williams (2010; 2014) has studied the linguistic practices of sixteenth and seventeenth century women, Joan and Maria Thynne, and Margaret Tudor. Similarly, Sairio (2006; 2014) has investigated the linguistic biography of Elizabeth Montague, a British social reformer in the eighteenth century, and Evans (2013) has examined the language of Queen Elizabeth I. In a forensic context, however, only Coulthard (1994) and Johnson and Wright (2014) focus on the idiolectal preferences of one individual.

It seems reasonable that, in author profiling, nomothetic (quantitative, generalised) and idiographic (qualitative, case study) approaches can be combined to garner and predict as much as possible about the author of a disputed document, both in terms of their social characteristics and what kind of communicants they are. Rather than relying on statistical results alone, quantitative evidence could act as a starting point in identifying linguistic features which are used with significantly different frequencies between groups of authors, for example males *vs* females. Following from this, an idiographic and qualitative approach could be applied, in terms of analysing *how* these features are used differently by different groups of people, as well as by individual authors within the groups. By combining qualitative and quantitative approaches at this early stage in the development of author profiling research, time and resources may be saved by avoiding replicating the situation in authorship attribution research (and criminal profiling) in which the two approaches have developed in competition with each other. Overall, the aim is to achieve as much linguistic evidence as possible, both quantitative and qualitative in nature, in

order to answer the question ‘what kind of linguistic person wrote this text?’ (Grant 2008: 223).

2.5 Chapter conclusion: addressing the issues

The overall contribution of this thesis to the field of forensic authorship analysis is that it aims to address the current related methodological and theoretical challenges in the fields of authorship attribution and authorship profiling detailed in this chapter.

First, the conceptualisation of idiolect has changed over recent years, especially within forensic linguistics. Today, a linguistic feature or pattern can be described as being distinctive of a person’s idiolect if they use it with higher consistency and frequency than is expected in the population from which they are taken. Chapter 5 sets out to further investigate this theory of idiolect with empirical evidence from the Enron Email Corpus. More specifically, research from corpus linguistics and psycholinguistics has theorised that collocations and the production of lexical sequences can be idiolectal and unique to individuals. The first of three analysis chapters in this thesis tests this theory, and aims to identify collocations and word n-grams for individual Enron employees which are distinctive at ‘population level’ when tested against the ‘Base Rate Knowledge’ for the corpus as a whole.

Second, research in authorship attribution has developed two diverging and competing approaches: the quantitative stylometric approaches and the more qualitative stylistic approaches. On the one hand, stylometric methods have the advantage of being more statistically reliable and objective than their stylistic counterparts. On the other, whereas the identification of style markers in stylistic approaches is grounded in a theory of linguistic variation, the linguistic features used and the statistical results produced by stylometric studies are difficult, if not impossible to explain or interpret in theoretical or linguistic terms. Therefore, Chapter 6 sets out to develop a methodology for authorship attribution which produces statistically reliable and replicable results using linguistic features that are underpinned by a theory of idiolect. Word n-grams are used in a series of attribution experiments which aim to identify the correct author of extracted and anonymised email samples.

Third, author profiling is a developing field in authorship analysis. The early work in this domain is focused on statistically correlating linguistic features with

social categories such as age, gender, nationality and personality. Such author profiling research is all stylometric, or nomothetic in nature. Author profiling faces the risk of finding itself in the same position as author attribution research, in that there are many quantitative and computational approaches that boast very good results, but offer little explanation as to the nature of the relationship between the linguistic features used and the social categories being explored. Chapter 7 focuses on using a corpus linguistic method to combine nomothetic and idiographic analyses in author profiling. These two approaches offer complementary types of evidence. The statistical analysis identifies linguistic features which discriminate between social groups in the Enron corpus. The stylistic analysis, drawing on aspects of collocation variation as in the previous chapters, provides more concrete and reliable evidence as to how these features are used by different groups, as well as by different authors within the same group.

The three analysis chapters use the Enron Email Corpus in different ways. The following chapter introduces Enron, the corpus, and the various sub-corpora and samples used throughout this thesis.

3 The Enron Email Corpus

3.1 Introducing Enron and the corpus

Enron is a former American corporation founded by Kenneth Lay in 1985 through the merger of two natural gas pipeline companies, Houston Natural Gas and Internorth. The newly merged company owned 37,000 miles of intra- and interstate pipelines for transporting natural gas between producers and utilities (Healy and Palepu 2003: 5). Through huge expansion and diversification, 'with a culture that was closer to Wall Street than a traditional utility company' (Stein and Pinto 2011: 703), Enron began extensively and internationally trading in natural gas, electric, water, coal, steel, weather derivatives, paper and pulp, broadband bandwidth and fibre optic cable capacity.

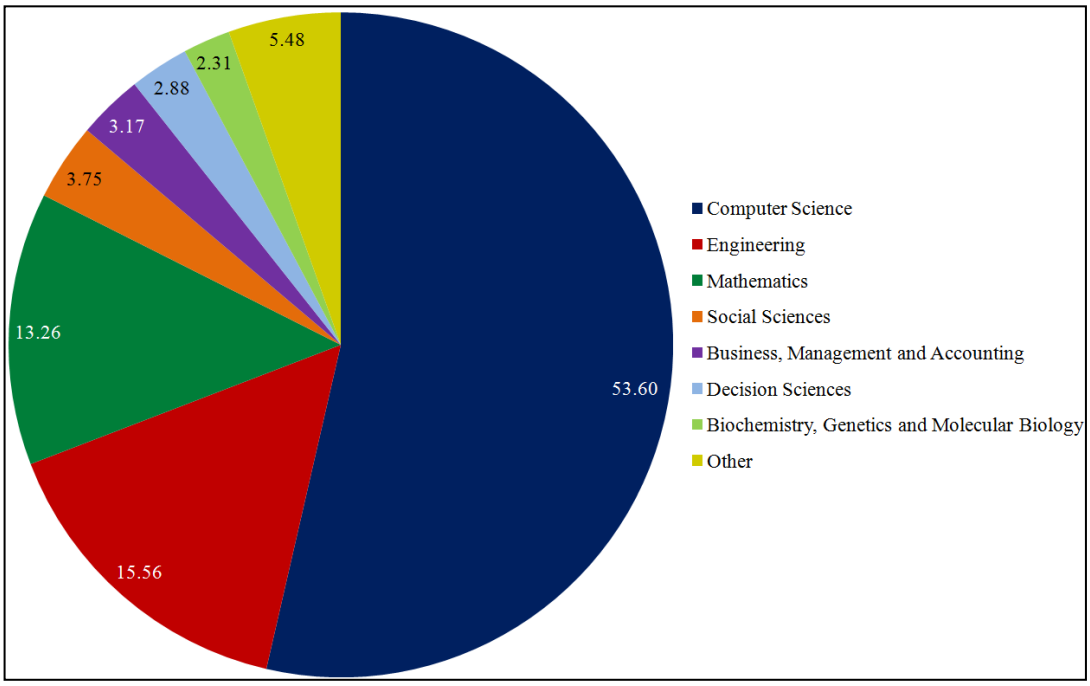
The success of Enron was startling. From the start of the 1990s until the end of 1998 Enron's stock rose by 311%, and this continued to soar, increasing by 56% in 1999 and 87% in 2000. By the end of 2001, Enron's stock was priced at $83.12, and its market capitalization exceeded $60 billion (Healy and Palepu 2003: 3). However, as a result of the sprawling nature of the company, Enron's business model was extremely complex, stretching the limits of accounting, and Enron 'took full advantage of accounting limitations in managing its earnings and balance sheet to portray a rosy picture of its performance' (Healy and Palepu 2003: 9). There exist a number of academic commentaries on the accounting practices of Enron (e.g. Benston and Hartgraves 2002; Powers et al. 2002). However, according to Healy and Palepu (2003), one of Enron's major financial activities was 'mark-to-market' accounting, which involves estimating the market value of long-term future contracts and (over)stating these projected profits as revenue on balance sheets. Another major factor in the accounting challenge for Enron was in the form of 'special purpose entities' or 'shell firms created by a sponsor, but funded by independent equity investors and debt financing' (Healy and Palepu 2003: 10). Enron designed several controversial special purpose entities specifically to fund the purchase of long-term contracts with gas suppliers, and even used them to hide the debt incurred from acquisitions. Essentially, Enron used special purpose entities primarily to achieve financial reporting objectives, allowing balance sheets to understate liabilities and overstate equity and earnings (Healy and Palepu 2003: 11).

As reports, accusations and allegations of accounting irregularities and malpractices emerged, the company's stock price was halved, and after a failed acquisition attempt by a smaller competitor, Dynergy, Enron filed for bankruptcy on 2 December 2001 with a stock price of only $0.26 (Healy and Palepu 2003: 11). News reports of the company's failing are available in *Business Week* (2001), *BBC* (2002) and *The Financial Times* (2002).

In 2003, as part of the Federal Energy Regulatory Commission's (FERC) legal investigation into the company's accounting malpractices, the email data of around 150 Enron employees, containing approximately 1.6 million emails sent and received between 1998 and 2002 was made publicly available online (FERC 2013). Since then, various versions of the data have emerged across the web (e.g. Wang 2009; Styler 2011; Fiore and Heer 2013; EDRM 2013; Illocution Inc. 2013). Many of these are modified versions of the set collected and prepared by Carnegie Mellon University (CMU) (Cohen 2009) in Pittsburgh, Pennsylvania, as part of its 'Cognitive Assistant that Learns and Organises' (CALO) project. It is this CMU dataset from which the corpora used in this thesis are extracted.

The Enron email dataset has been widely used for research purposes across an expansive range of fields. Cohen (2009) lists some early studies which use the Enron data for research into email data management and classification (Klimt and Yang 2005; Bekkerman et al. 2005) and social network analysis (McCallum et al. 2007). As of August 2014, a search for the key terms *Enron* and *email* in SciVerse's Scopus—'the largest abstract and citation database of peer-reviewed literature' (SciVerse Scopus 2013)—returns 229 results for conference papers and journal articles since 2004. These studies come from 15 different fields (Figure 2), with the majority being in computer science (53.6%), and 15.56% and 13.26% from engineering and mathematics respectively. There is substantial overlap in the type of work emerging from these three inter-related fields. In line with the work noted by Cohen (2009) many modern studies of computer science, engineering and mathematics relate to social network analysis (Lubarski and Morzy 2012; Gliwa et al. 2012; Dikmen and Huang 2011), and email classification (Li et al. 2012; Tam et al. 2012; Al Sallab and Rashwan 2012).

Figure 2. Fields of research results of a Scopus search for *enron* and *email*



Further, the Enron data has been used for authorship analysis research under a variety of different methodological guises, including text categorisation (Wang et al. 2010), data mining (Geng et al. 2008), authorship disambiguation (Maes and Scholtes 2012), authorship verification (Brocardo et al. 2013), and authorship similarity (Iqbal et al. 2008 and 2010; Neumann and Schnurrenberger 2009; Chen et al. 2011). Despite this promising amount of authorship work being undertaken using the Enron data, it is being produced almost exclusively by computationalists. Linguists *have* worked with the Enron data, but far less frequently, and their research does not relate to authorship analysis or forensic linguistics. For example, Lampert et al. (2008) focus on requests and commitments in the Enron emails. Using Speech Act Theory (Austin 1962; Searle 1969), their paper offers precise definitions for automatically classifying requests and commitments in emails, with the ultimate aim being to train an email system or client to identify utterances in emails which place responsibility for action on the users themselves or others, and manage email accordingly. Kessler (2010) takes a very traditional corpus linguistic approach and explores the contemporary meaning of the word *virtual* in the Enron corpus. The data is useful for this purpose, as Enron was 'a tech savvy company [...] referred to by some as a virtual company' (Kessler 2010: 262). Gilbert (2012) uses the Enron corpus to explore the relationship between the words people use and their rank in the corporation, identifying phrases which signal workplace hierarchy. Mitra

and Gilbert (2012) in the analysis of workplace gossip, investigate at which levels of the corporation it is most common and who is most responsible for circulating it. Most recently, Titak and Roberson (2013) replicate Biber's (1988) corpus-based multi-dimensional analysis in investigating the structural and functional linguistic variation in online language across a range of web registers, and Enron emails make up their 'workplace email' register. In general, though, computational approaches to the Enron email dataset far outweigh the linguistic.

This thesis, Wright (2013) and Johnson and Wright (2014) are the first studies to utilise the corpus within the field of forensic linguistics, and the data holds a number of advantages for forensic authorship analysis in particular. First, emails are an enduring medium of online communication, flourishing even in the social media age of blogging, Facebook and Twitter. A report by technology market research firm Radicati Group Inc. (2013) found that there is a total of 3.9 billion email accounts registered worldwide in 2013, a number expected to grow to 4.9 billion by 2017. It also reported that 182.9 billion emails were sent and received daily worldwide in 2013 which, by 2017, is expected to rise to 206.6 billion. In particular, the report (2013: 2–3) identifies email as 'the predominant form of communication in the business space' with 929 million business email accounts registered, accounting for over 100 billion sent and received emails per day. As emails continue to be a major way in which we communicate on the web, there are also increasing instances in which they are misused, as Coulthard et al. (2011: 538) comment: 'growingly, such [forensic] cases involve email' which contain 'threatening, abusive or defamatory material'. Given emails are increasingly becoming central to forensic cases, there is a growing demand for authorship analysis research which focuses on this particular text type. There has been such research published and presented by forensic linguists (e.g. Turell 2010; Coulthard 2013; Grieve 2013), but this has largely been limited to the reporting of casework. As such, the Enron corpus represents a unique opportunity for empirical research with implications for authorship analysis casework.

As Cohen (2009) points out, the Enron dataset is 'the only substantial collection of "real" email that is public'. The fact that this data is naturally occurring has benefits for corpus linguistics generally and authorship analysis specifically. This concept of naturally occurring or 'real' language, stays true to the motivation that fuelled the inception of corpus linguistics, as Leech (1992: 105) states:

> For many older linguists, the term ‘corpus linguistics’ is evocative of the heyday of the corpus in the 1950s: the era of Harris, Fries, Hill and other American structuralists, for whom a corpus of authentically occurring discourse was *the* thing that the linguist was meant to be studying (original emphasis).

The Enron employees at the time, by and large, would not be able to predict the future demise of the company or the release of the emails, and even less so the fact that their emails would be used for linguistic study. As such, the data is far removed from the influence of the ‘observers’ paradox’ (Labov 1972: 209) and offers a greater authenticity than texts contrived or manufactured for the sole purpose of research. An obvious but important advantage for authorship work is that with copied and pasted material removed (see Section 3.2) we can be confident in the knowledge that the individual whose email address an email is attributed to is the sole ‘executive author’ (Love 2002: 43) of the language within that email. That is, we can identify the individual who has ‘engaged in formulating the expression of ideas and made word selections to produce the text’ (Grant 2008: 218). Finally, the email texts are already in digital form, avoiding the time-consuming and arduous task of digitising hard copies of texts to make them amenable to corpus-assisted analysis (Cotterill 2010: 580). That said, as de Vel et al. (2001: 59) comment, when working with publicly available email data it is ‘generally quite difficult to obtain a sufficiently large and clean corpus’, with ‘clean’ being glossed as ‘void of cross-postings, off-the-topic spam, empty bodied emails with attachments etc.’. This is the case with the Enron data as made available by Cohen (2009) and others, and as such the data was extracted and prepared in such a way that it is suitable for authorship analysis, the process of which is detailed in the following section.

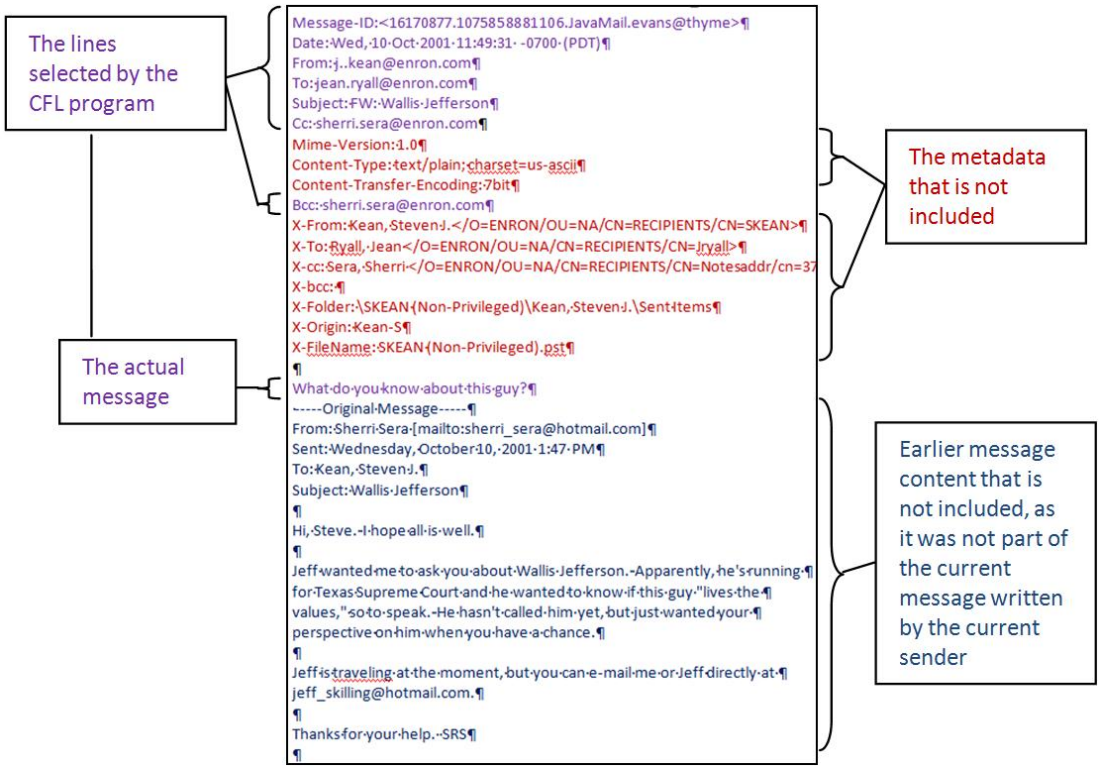
3.2 Extracting and preparing the data

The full Enron corpus was downloaded from https://www.cs.cmu.edu/~enron/ and the subsequent extraction and ‘clean-up’ of the data was performed by David Woolls, CEO of CFL Software Ltd, using specialist extraction software. A technical report of the extraction procedure has been provided by Woolls and is given in Appendix 1.

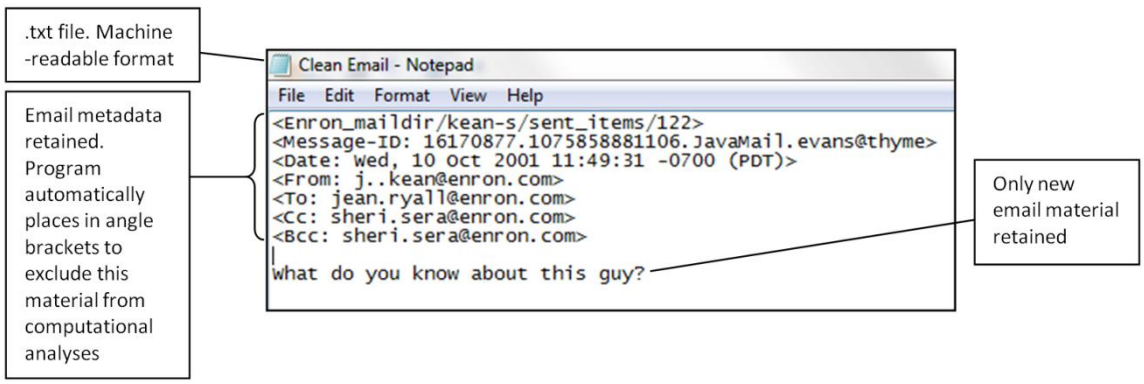
In the CMU dataset, each email file contains the entire thread of email conversation that has taken place prior to the email in question, from the original email up to the most recent message (Figure 3).

Figure 3. Comparing an original CMU version email with the cleaned up version of the type used in this study

CMU Version



Cleaned up version



This is problematic for authorship analysis, as not all of the language in such email files belongs to the author whose sent box the email is in. Therefore, in the extraction process, all previous emails in a thread were removed, and only the newest message, sent by the current sender, was retained. This way we can be sure that the individual in question is the author of the email; in other words, the language found in each email file can be confidently attributed to the correct sender (with the exception of assistants writing on behalf of their superiors, discussed below). In addition, in the CMU version each email was accompanied by a substantial amount of metadata, much of which was not required for this study. As such, the only metadata retained with the emails in this study are the date and time the email was sent, along with the 'From:', 'To:', 'Subject:', 'Cc:' and 'Bcc:' fields (Figure 3). This metadata was retained so that any linguistic variation observed in emails throughout the analyses in this thesis can be considered in relation to sociolinguistic and contextual features such as number and type of recipient(s) and the subject of the email. The metadata included with each email was enclosed in angle brackets (< >) so that it is excluded from computational analyses. The corpus is organised and saved with all of the emails belonging to the same author being aggregated in one plain text (.txt) file. Having an individual file for each author allows for authors' data to be compared and contrasted in various ways, including comparing the linguistic patterns of one author with the rest of the authors in the corpus (Chapter 5 and 6), or grouping authors together on the basis of social variables such as gender or occupation (Chapter 7). Plain text format was used so that the data was compatible with a range of computer software packages.

The collection of the Pilot Corpus (Section 3.4.2), consisting of only four authors, provided an opportunity to test the automated extraction process, identify any limitations of it and rectify these before extracting the whole corpus. Indeed, the automatic extraction and cleaning up of the CMU Enron data presented a number of challenges. One problem was that of empty emails, which did not include any language at all and were presumably sent only for the transfer of attachments. Such instances were removed from the Pilot Corpus sample and the extraction software was coded to exclude such emails from the collection of the whole corpus. Similarly, because the sent emails of the Enron employees were distributed across different sent folders (e.g. 'sent', 'sent_items', 'sent_mail') there were often duplicate emails found, where the same email appeared in more than one sent folder.

This was resolved by identifying identical information in the metadata fields. The initial thought was that the Message ID field could have been used to identify duplication of emails. However, strangely, email Message IDs provided in the CMU version were not unique. Instead, the 'Date:' field was used as a indicator of duplication, and those emails which were sent within the same second were flagged as being duplicated. This proved a reliable method for the extraction program to automatically identify and remove duplicate emails, so that the email in question appeared only once. This duplication issue was spotted during the pilot phase, so the solution was implemented in the extraction of the rest of the corpus. There were also instances in which the emails within the sent items of a particular employee were not sent from that person's email address. This was mostly the case for high-ranking employees who had personal assistants who sent emails on their behalf. Emails of this kind were sent from the assistants' email address, but saved in their bosses' sent folders. Such emails were relatively easy to identify, as the email address in the 'From:' field was that of the assistant. Once identified, these emails were removed. In cases in which there was more than one email sent from the same assistant, these emails were extracted and saved as belonging to this assistant, essentially creating a separate set of files for this individual, and treating them as an additional author. A total of 23 authors were added to the corpus in this way. Although these emails appear to be sent by the assistants, it is impossible to be sure that the content of the messages has not been dictated to the assistant by their boss, which was a challenge Coulthard (2013: 459) encountered in an email authorship case.

There was the potentially confusing issue that some authors had two different email addresses in the Enron corpus, which, according to Priebe et al. (2005: 5), is a result of some post-processing. For example, the author Kevin M. Presto has two email addresses. The first and most common in the corpus is that of first name and last name, such as in kevin.presto@enron.com. However, Presto and some other authors have email addresses made up of the initial of their middle name and their full last name, such as m..presto@enron.com. Such cases were relatively straightforward to deal with, and the emails from both email addresses for the same author were grouped together in the same file. There were also more problematic cases, however, where two different authors had the same last name and same first initial, and so their emails had been grouped together in the CMU set. For example,

the emails of Jeff Hodge and John Hodge were both in the 'j-hodge' sent folders. In such instances, the same procedure was used to identify which of the two employees the email belonged to using the 'From:' field and the first name as shown in the full Enron email addresses. The emails of the two authors were separated and saved in different files. This issue arose with the email sets of five of the authors in the CMU set, each of which were subsequently divided into two different authors. Finally, there were emails which comprised, partially or entirely, text copied and pasted from a third party, most often from online news reports, or emails which they had received from someone else and were forwarding on. Again, identifying such emails was relatively straightforward, as they were often markedly longer than average, and noticeably started with the journalist's name and newspaper, or signed off by someone from another company. Those emails which were entirely copied from elsewhere were removed manually from the dataset, and if the emails were only partly copied from elsewhere, then the copied section was removed manually. Overall, this extraction and clean-up process ensured that the data used in this study is as reliable and accurate as possible before undertaking any analysis. The extraction process used and described here was developed specifically for dealing with the CMU version of the Enron corpus. It is important to note that in forensic casework involving email evidence, the preparation of the data for analysis will depend on the nature and form in which the linguist receives the data, either from the police or legal parties. The email client used (e.g. Gmail, Outlook, or a corporate one used within a company) will influence how accessible the data is in the first instance. Furthermore, it is likely that in each case, the type and format of metadata held with each email and the way in which old and new material is demarcated in an email thread will vary. The implication for this is that different datasets in different cases will require different extraction and clean-up procedures, and there is not a 'one-size-fits-all' technique for preparing email data. That said, datasets may not be as large or 'dirty' as the Enron one, and once the conditions on which emails can be separated and metadata retained have been identified, the actual process of programmatical extraction is not a time consuming one.

3.3 Ethical considerations

As reported by the *Wall Street Journal* (Berman 2003), shortly after FERC released the email documents, Enron petitioned to get some of the most private and sensitive

messages removed. As a result, the email database was disabled for ten days while Enron and its employees identified every email that they wanted to be removed from the public eye. FERC eventually removed about 8% of the database, or 141,379 documents (Berman 2003). The result of this is that those emails which remain in the public domain do so with the permission of Enron and the employees to whom the emails belong, and are therefore publicly available for research purposes. Cohen (2009), in his distribution of the email database online warns, 'in using this dataset, please be sensitive to the privacy of the people involved', and continues: 'many of these people were certainly not involved in any of the actions which precipitated the investigation'. The purpose of giving examples or extracts of emails in this thesis is not to incriminate the authors in any way, or to air any private correspondence. Rather, the focus here is entirely on a particular linguistic feature(s) that the author uses in their emails. Emails have not been anonymised here. Because the content of the emails can easily be included in an internet search, and the senders of emails readily identified, any anonymisation of data or extracts used in thesis would be ineffective. Sensitivity has been exercised in that personally identifiable information such as telephone numbers, addresses, National Security numbers, bank details and software or website passwords are not included in extracts or examples at any point throughout this thesis. In addition to this, any emails which include private, sensitive or potentially offensive or upsetting content, such as jokes or details about the loss or illness of family members, are not used as examples in this study. The Enron data, and the form and purposes with which it is used in this study, have passed the University of Leeds' Arts and PVAC Faculty Research Ethics Committee's Light Touch Ethical Review, the decision of which is included as Appendix 2.

3.4 Breakdown of the corpus

The Enron Email Corpus is used in various ways throughout this thesis: as a reference corpus (Chapter 5), as a pool of candidate authors (Chapter 6) and as being representative of the linguistic practices of various groups of employees in the corporation (Chapter 7). In order to use the corpus in these ways, as well utilising the corpus as a whole, various samples have been constructed.

3.4.1 The full Enron Email Corpus (EEC)

Following the extraction and preparation procedures described above (Section 3.2) the resulting corpus used in this thesis comprises 176 authors, 63,369 emails and 2,462,151 tokens (Table 1).

The full Enron Email Corpus can be used as a reference corpus against which the frequency or rarity of particular lexical items and lexical strings observed in the data of individual writers can be measured, to identify how distinctive they are of these writers. The Enron Email Corpus is a different type of reference or comparison corpus to those large general English corpora such as the *British National Corpus*, (c.100 million words) or the much larger *Collins Corpus* (c.2.5 billion words). Rather, it is a specialist corpus, compiled—or in this case prepared—for a specific purpose, within a particular setting and with particular participants and of a specific text type (Flowerdew 2004: 21). As McEnery et al. (2006: 15) note, whereas 'representativeness' of a general language corpus typically refers to the extent to which it can serve as a basis for an overall description of language or language variety, specialised corpora tend to represent a domain (e.g. medicine or law) or a genre. In the case of the Enron corpus, its representativeness relates to the Enron population; it represents the email interactions of the Enron 'community of practice', 'an aggregate of people who come together around mutual engagement in some common endeavour' (Eckert and McConnell-Ginet 1998: 490). Therefore, while some authors, such as lawyer Sara Shackleton have large amounts of data (3,465 emails, 148,518 tokens), and others, such as administrative assistant Joseph Alamo (2 emails, 245 tokens) have very little, the corpus represents all of the data available in the various sent folders for all 176 authors.

Table 1. Breakdown of the Enron Email Corpus (EEC)

Author	Emails	Tokens	Author	Emails	Tokens
Akin, Lysa	11	320	Gay, Randall	111	3,777
Alamo, Joseph	2	245	Gay, Rob	119	5,541
Allen, Phillip	360	16,710	Geaccone, Tracy	393	7,449
Arnold, John	1,040	26,659	Germany, Chris	2,341	91,621
Arora, Harry	50	1,752	Gilbertsmith, Doug	13	664
Badeer, Robert	35	1,036	Giron, Darron	452	15,415
Bailey, Susan	11	605	Griffith, John	25	1,371
Bass, Eric	1,009	25,481	Grigsby, Mike	478	19,277
Baughman, Don	39	3,602	Guzman, Mark	246	13,227
Beck, Sally	1,272	104,679	Haedicke, Mark	804	19,622
Benson, Robert	7	519	Hain, Mary	146	10,288
Blair, Lynn	804	22,040	Hayslett, Rod	530	9,398
Brawner, Sandra	90	3,181	Heard, Marie	613	24,269
Brown, Katherine	5	116	Helton, Jenny	3	32
Brown, Kimberly	3	122	Hendrickson, Scott	17	694
Buy, Rick	481	14,817	Hernandez, Juan	76	4,647
Campbell, Larry	226	28,482	Hernandez, Judy	18	548
Carson, Mike	70	1,782	Hillis, Kim	43	1,799
Cash, Michelle	889	31,419	Hodge, Jeff	79	1,909
Causholli, Monika	181	7,161	Hodge, John	15	519
Chapman, Kay	24	1,138	Holst, Keith	24	2,606
Corman, Shelley	405	15,733	Horton, Stanley	271	9,389
Crandell, Sean	97	39,443	Hyatt, Kevin	170	8,546
Cuilla, Martin	63	2,560	Hyvl, Dan	474	26,708
Dasovich, Jeff	2,855	170,316	Jones, Tana	2,990	123,231
Davis, Dana	77	2,636	Kaminski, Vince	2,306	54,498
Dean, Clint	4	261	Kean, Steven	1,118	33,751
Dean, Craig	8	311	Keavey, Peter	55	2,064
Delainey, David	677	26,778	Keiser, Kam	327	13,108
Derrick, James	470	6,042	King, Jeff	8	155
Dickson, Stacy	145	3,937	Kitchen, Louise	911	25,899
Donoho, Lindy	180	7,520	Kuykendall, Tori	113	4,735
Donohoe, Tom	15	885	Lavorato, John	1,114	25,320
Dorland, Chris	502	14,605	Lay, Kenneth	9	738
Elbertson, Janette	47	1,953	Lenhart, Matthew	1,059	21,064
Ellis, Kaye	40	868	Lewis, Andrew	23	766
Ermis, Frank	15	260	Linder, Eric	6	124
Farmer, Daren	772	24,502	Lokay, Michelle	163	6,946
Fischer, Mary	45	1,387	Lokey, Teb	90	2,095
Fisher, Mark	19	1,250	Love, Phillip	981	36,471
Fleming, Rosalee	88	2,427	Lucci, Paul	69	2,150
Forney, John	333	21,315	Maggi, Mike	117	939
Fossum, Drew	876	39,297	Mann, Kay	3,055	103,512
Gang, Lisa	56	1,662	Martin, Tom	169	8,516

Table 1. (cont.) Breakdown of the Enron Email Corpus (EEC)

Author	Emails	Tokens	Author	Emails	Tokens
May, Larry	39	635	Sera, Sherri	81	3,463
McCarty, Danny	141	6,183	Shackleton, Sara	3,465	148,518
McConnell, Mark	62	2,389	Shankman, Jeffrey	939	20,920
McConnell, Mike	613	34,188	Shapiro, Richard	7	286
McCulloch, Angela	4	107	Shively, Hunter	265	5,730
McKay, Brad	27	437	Skilling, Jeff	29	1,292
McKay, Jonathan	193	7,045	Slinger, Ryan	28	1,631
McLaughlin, Errol	323	8,700	Smith, Gretel	2	20
McVicker, Maureen	114	4,507	Smith, Matt	318	14,631
Merris, Steven	2	68	Solberg, Geir	37	1,535
Meyers, Albert	7	530	South, Steven	6	108
Mims-Thurston, Patrice	241	13,779	Staab, Theresa	68	2,811
Neal, Scott	404	8,773	Stark, Cindy	40	1,426
Nemec, Gerald	1,475	58,911	StClair, Carol	990	55,147
Panus, Stephanie	16	598	Steffes, Jim	1,202	35,668
Parks, Joe	213	4,498	Stepenovitch, Joe	52	1,919
Pereira, Susan	65	2,141	Stephens, Beverly	7	472
Perlingiere, Debra	1,634	59,149	Storey, Geoff	53	1,441
Phillips, Cathy	32	1,938	Sturm, Fletcher	235	8,883
Pimenov, Vladi	38	940	Sweet, Twanda	100	2,641
Platter, Phillip	43	1,804	Swerzbin, Mike	43	1,586
Presto, Kevin	816	23,077	Symes, Kate	1227	58,754
Quenet, Joe	43	825	Taylor, Liz	85	2,366
Quigley, Dutch	325	9,882	Taylor, Mark	1548	71,849
Rangel, Ina	3	218	Tholt, Jane	268	8,904
Rapp, Bill	88	3,771	Thomas, Paul	100	2,856
Reitmeyer, Jay	30	1,138	Thompson, Patti	54	2,894
Richey, Cooper	137	4,833	Townsend, Judy	10	191
Ring, Andrea	67	3,143	Tycholiz, Barry	395	16,356
Ring, Richard	33	3,135	Ward, Kim	509	20,262
Rodrigue, Robin	643	28,208	Watson, Kimberly	797	23,007
Rogers, Benjamin	616	20,552	Weldon, Charles	231	9,720
Ruscitti, Kevin	91	2,980	Wells, Tori	3	76
Sager, Elizabeth	791	23,993	Whalley, Greg	139	3,861
Saibi, Eric	11	264	White, Stacey	370	12,384
Salisbury, Holden	95	2,929	Whitehead, Brenda	7	661
Sanchez, Monique	28	1,306	Whitt, Mark	151	3,903
Sanders, Richard	959	17,466	Williams, Bill	424	24,183
Sauseda, Sylvia	2	41	Williams, Jason	29	665
Scholtes, Diana	56	1,871	Williamson, Joannie	17	329
Schoolcraft, Darrell	264	7,475	Wolfe, Jason	56	2,106
Schwieger, Jim	73	4,880	Y'Barbi, Paul	94	4,397
Scott, Susan	880	61,482	Zipper, Andy	247	7,912
Semperger, Cara	148	4,169	Zufferli, John	214	6,217

It is well documented that the amount of population data is very limited for forensic linguists, particularly when compared with what is available for DNA analysts and forensic phoneticians (Coulthard 1994; 2010; Grant 2007; 23; Koehler 2013: 516). Coulthard (2013: 466) goes as far as to say that 'forensic linguists are never going to have reliable population statistics to enable them to talk about the frequency or rarity of particular linguistic features'. However, the importance of *relevant* comparative reference corpora in the identification of distinctive language use was discussed above (Section 2.1), with forensic linguists now turning to smaller more specialised comparison corpora. In the context of forensic case work, Butters (2012: 532) highlights that one of the 'Aims' of the International Association of Forensic Linguists (IAFL) is: 'collecting a computer corpus of statements, confessions, suicide notes, police language, etc., which could be used in comparative analysis of disputed texts' (IAFL 2006). Similarly, McMenamin (2004: 78) argues that one 'very promising area of linguistic research is the use of large corpora to provide 'baseline data for determining the relative frequency of occurrence (i.e. identifying or discriminating potential) of specific style markers used in a given population of writers'. McMenamin (2010: 504), when discussing reference corpora states that 'the corpus for a given case should match as much as possible the context of writing of the text(s) under scrutiny'. In addition, Coulthard (2010: 483) says that for reference and comparative purposes, the linguist must establish what a 'relevant population of speakers or of language samples' is for each case. Finally, Butters (2012: 354) in his questions and suggestions on standards and best practices for forensic linguistics asks:

> Can we not require that the validity of every putative marker that we introduce into evidence in authorship analysis be attested to in some significant way, *such as comparison to a sociolinguistically valid or a statistically meaningful comparison corpus?* (My emphasis).

Although smaller than general reference corpora, the Enron Email Corpus is a sociolingusitically and statistically valid and relevant reference corpus for testing the distinctiveness of linguistic features found in the emails of individual Enron employees. Turell and Gavaldà (2013: 499) define Base Rate Knowledge as the usage of linguistic features 'by a relevant population, or group of language users from the same linguistic community'. The Enron Email Corpus satisfies these criteria, and so provides a Base Rate Knowledge for language use in the Enron

corporation. For example, if a particular linguistic feature or pattern is found to be common in the emails of one employee, the frequency of this feature or pattern can be checked in the Enron Email Corpus as a whole,. This way, we can identify how common it is within the Enron population, the inhabitants of which are writing in the same medium from within the same company and at the same time. This is a more relevant and specific comparison than testing the frequency of this feature in a much larger reference corpus, which covers many fields, topics, varieties of English, dates, and text types, or using the web as a reference corpus. Therefore, the Enron Email Corpus offers a suitable, relevant and effective reference corpus for comparing emails written by any given Enron employee(s) with emails written by other Enron employees, providing important and useful normative data produced within very similar contexts of writing.

The size of the Enron Email Corpus as a dataset for authorship research is favourable. Grant's (2004; 2007) General Authorship Corpus, for example, made up of a variety of different text types ranging from short stories to personal letters, comprised 50 authors, 175 texts and 124,435 words. Similarly, the corpus is as large, or larger, than other specialist reference corpora being compiled by (forensic) linguists. Of course, this is likely to do with the accessibility of the texts in question; emails are far easier to collect than suicide notes, for example. The corpus is larger than the 338 text corpus of Colorado writers that McMenamin used as a comparison corpus in the JonBenét Ramsey case (McMenamin 2002: 197) and the impressive 286-text suicide note corpus collected by Shapero (2011). It also contains more texts than the 756-message and 11,067-message SMS corpora collected by Dyer (2008) and Tagg (2009) respectively. That said, although it is likely to be larger in terms of tokens, the Enron corpus contains fewer authors than the 1,100 author American letter writer corpus used in casework and cited by McMenamin (2010: 489). Similarly, though the Enron corpus contains more texts than the SMS corpus provided to Grant (2010) by Northamptonshire police, it contains slightly fewer that the 186 authors Grant's corpus included. The Enron Email Corpus is also comparable in size to some of the very large sets of data used by computationalists, such as Savoy (2012) (c. 3.5 million words) and Argamon and Levitan (2005) (2 million words). However, it is far smaller than other computational authorship corpora, such as that used by Koppel et al. (2011: 85) which included 2,000 words from 10,000 blogs equalling 20 million words. Similarly, Narayanan et al. (2012)

ran algorithmic tests for authorship on a dataset of 2,443,808 blog posts, with an average of 305 words, giving an approximate corpus size of 745,361,440 tokens. Given that one of the methodological aims of this paper is to combine stylistic and statistical evidence, the specialised nature of the corpus has the advantage over these much larger datasets in that 'they allow a much closer link between the corpus and the contexts in which the texts in the corpus were produced' and so 'quantitative findings revealed by corpus analysis can be balanced and complemented with qualitative findings' (Koester 2010a: 67). Finally, it is large enough to make meaningful statistical comparisons between one author and the rest of the corpus. The author with the most tokens is Jeff Dasovich with 170,316 tokens. His set of emails accounts for only 6.92% of all the tokens in the corpus. Therefore, as a reference corpus it is 14 times larger than the largest individual set within it. Berber-Sardinha (2000: 12) claims that in the identification of statistically salient linguistic items 'a reference corpus does not need to be more than five times larger than the study corpus'.

Overall, the Enron Email Corpus in its entirety serves to offer normative population baseline data for the identification of author- or group- distinctive linguistic features. In addition to this, however, in order to pursue specific research aims throughout the chapters of this thesis, it is broken up into various, smaller sub-corpora: the Pilot Corpus, the twelve-author sample (EEC12) and the eighty-author sample (EEC80).

3.4.2 The Pilot Corpus

Prior to the full authorship attribution experiment reported in Chapter 6, a small scale pilot study was undertaken to test the methodology (detailed in Section 4.5 below), using a corpus of four authors, 2,622 emails and 86,902 tokens (Table 2).

Table 2. Breakdown of The Pilot Corpus

	Arnold	**Germany**	**Lavorato**	**Zipper**	**Total**
Emails	632	1,339	405	246	2,622
Words (tokens)	20,890	47,543	9,721	8,748	86,902
Words (types)	3,026	3,788	1,857	1909	10,580
Average email length (tokens)	33	36	24	36	33

Four authors were chosen as it is a number which facilitated enough comparison across authors to evaluate the method at the initial stage, while still being a number small and manageable enough so that the data could be read and the data clean-up procedure (see Section 3.2) could be developed and fine-tuned. This Pilot Corpus accounts for 2.27% of the 176 authors in the full corpus, 4.14% of the total emails and 3.53% of tokens. It was during the extraction and preparation of this pilot corpus that a number of issues were identified and remedied before the extraction of the full corpus (detailed in Section 3.2) from the CMU (Cohen 2009) version.

The only decisions made in the compilation of the corpus for the pilot study were in relation to the gender and occupation of the authors, and the size of their datasets. The aim was to select four writers who are as socially similar as possible. As such, John Arnold, Chris Germany, John Lavorato and Andy Zipper are all male, they are all traders, they are all American, they were all of working age at the time of writing and, of course, they are all using the same mode of communication (email). The overriding similarity between these four authors is that, although they spread across various hierarchical levels of the corporation, their role within the company was in trading, that is, their jobs were to buy and sell energy and commodities in Enron's online market place on a daily basis. Traders were chosen for the pilot because they were the group of employees that was initially available for analysis; they were the first to be collected, extracted and processed by Woolls. In addition, their high levels of register-specific language made them easily identifiable as a distinct group of employees who shared related roles, and this posed an interesting first authorship challenge for the method. In the various sources that were used to determine the authors' occupations (Baucus and Grassley 2003; Creamer et al. 2009; Priebe et al. 2010), Germany and Zipper are referred to as 'traders'. Arnold and Lavorato are also traders, but occupy higher positions in the company's hierarchy; Arnold was also a Vice President of the company, and Lavorato was a President. This difference in hierarchical rank of these four authors could affect their language use (Chapter 7 examines occupation and language), as some emails will be more 'managerial' than 'trader' in nature. However, unlike in some computational authorship studies (e.g. de Vel et al. 2001; Savoy 2013), there were no measures taken to control or code emails for topic. Given that these four authors are all involved in trading in some way, though, the hypothesis was that many of their emails were likely to have topics and subject matter in common, and

that they draw upon much of the same trader-related 'register' (Halliday and Hasan 1985: 12). Therefore, the assumption was that they will make use of similar linguistic features. Following Kredens (2002: 406), the rationale behind choosing such socially similar writers is that if distinctive variation can be identified between these authors in the pilot, then such variation will be even greater between writers of dissimilar social characteristics, and this indicates that the method has good potential for attributing authorship.

Once the authors for the Pilot Corpus had been decided upon, all of the emails were extracted from the single largest sent folder for each author. The sent folders of some authors in the Enron dataset contain more emails than others, so this resulted in there being an imbalance in the amount of data available for the four authors. For example, Germany has more than twice as many emails as Arnold, the author with the second largest sent folder. This imbalance was left unaltered, and the amount of data used for each author was not normalised in any way. This decision was taken as this reflects the reality of how much data the police would obtain for each of these authors if they seized their most active sent folder. It also avoids difficult sampling issues (e.g. Grant 2007). Most importantly, however, it allows for the analyst to examine the extent to which the methods employed respond to the challenges posed by varying dataset sizes for different authors.

3.4.3 The twelve-author sample (EEC12)

Chapters 5 and 6 make use of a twelve-author sample of the Enron Email Corpus (henceforth EEC12), comprising 12,633 emails and 382,070 tokens (Table 3). EEC12 is an expansion of the Pilot Corpus, adding eight authors to the four used in the pilot, and containing 4.82 times as many emails and 4.4 times as many tokens. In EEC12, the emails used for each author are taken from *all* of the various sent folders for that author, rather than only the largest one as in the Pilot Corpus. Overall it accounts for 6.82% of all 176 employees in the Enron Corpus, 19.94% of all emails and 15.52% of all tokens.

Table 3. Breakdown of the EEC12 sample

Author	Emails	Tokens	Mean email length (words)
John Arnold	1,040	26,659	25.6
Chris Germany	2,341	91,621	39.1
John Lavorato	1,114	25,320	22.7
Andy Zipper	247	7,912	32.0
Phillip Allen	360	16,710	46.4
Chris Dorland	502	14,605	29.1
Daren Farmer	772	24,502	31.7
Vince Kaminski	2,306	54,498	23.6
James Derrick	470	6,042	12.9
Mark Haedicke	804	19,622	24.4
Gerald Nemec	1,475	58,911	39.9
Jim Steffes	1,202	35,668	29.7
Total	**12,633**	**382,070**	**29.8**

Chapters 5 and 6 are interested in identifying individual author-distinctive linguistic variation which may or may not provide empirical evidence of their idiolects. Therefore, the important aspect of the EEC12 sample is that, by using all of the emails from all their sent items, it represents their linguistic (email) activity as fully as possible. Essentially, any twelve of the full Enron Email Corpus could have been used, but the aim when building the EEC12 sample was to include authors from range of different occupations in the corporation. The sex of the authors remained consistent, with all twelve being men. In addition to the four employees involved in Enron trading, four authors were added whose jobs were within legal departments. James Derrick was Enron's chief in-house lawyer and General Counsel, and Gerald Nemec was attorney for Enron Capital and Trade Resources. Mark Haedicke was a managing director as well as being General Counsel, and Jim Steffes worked in governmental legislation and was Vice President. In addition to these, Chris Dorland and Daren Farmer were identified as 'managers' in Enron, while Vince Kaminski was managing director for research and Phillip Allen was also managing director. These four managers are different from the traders and the lawyers insofar as their job titles and roles as documented in Baucus and Grassley

(2003), Creamer et al. (2009) and Priebe et al. (2010) explicitly identify them as being ‘managers’ or ‘managing directors’ *and* as having no reported relationship with either trading or legal departments. This contrasts with Arnold and Lavorato who were both identified as being ‘traders’ *and* holding (vice) presidential roles, and Haedicke who was a managing director but also general council. Overall, while the analysis using the EEC12 corpus is concerned with individual authors rather than groups sharing the same occupation, three general kinds of employees are covered in this sub-corpus: traders, lawyers and managers.

As with The Pilot Corpus, emails were not explicitly coded for topic (besides the ‘Subject’ field in the metadata) and no normalisation of size of individual author’s datasets took place. The only criterion with regard to size was that the authors had no fewer emails than Andy Zipper (247). This results in some authors (Germany, Kaminski, Nemec) having a great deal of data compared with others (Derrick, Zipper). Burrows (2007: 30) states that datasets of around 10,000 tokens ‘suffice as a reliable minimum for an authorial set’ in authorship attribution research. However, as Luyckx and Daelemans (2011: 38) highlight in relation to the amount of data used per author in authorship tests, ‘the effect of data size has not been researched in much detail yet, since most stylometry research tends to focus on long texts per author’. Indeed, studies such as Argamon and Levitan (2005), Labbé and Labbé (2001), Burrows (2002; 2003; 2007), and Savoy (2012), amongst others, use hundreds of thousands of words per author—mainly literary texts—in their research. The amount of data per author in EEC12 is far less than that used in such studies, with the smaller sets dropping below the 10,000 token threshold. The substantial variation in the amount of data available for each author presents useful methodological challenges in the thesis. Having a dataset with the distribution of EEC12 may highlight the benefits and drawbacks of particular analytical approaches used throughout this thesis with regard to dataset size. The difference in dataset size across authors is to be expected, given that people produce different levels of linguistic output, via different modes of communication, both inside and outside of the workplace, and therefore reflects individual email volume and activity.

EEC12 is used in two different ways in Chapter 5 and 6. In Chapter 5, it is used as a sample for the corpus stylistic comparison of the twelve authors in terms of their lexico-grammatical patterns and preferences when using very common lexical items. In the first instance, the twelve authors are compared with each other,

and when potentially author-distinctive lexico-grammatical preferences, or word n-grams, emerge from this initial comparison, their frequency of occurrence is tested against the full Enron Email Corpus as a reference set. Using this approach, it is possible to identify distinctive lexico-grammatical choices made by individual writers in the corporation, which are able to distinguish them from the Enron population from which they are taken. In Chapter 6, the twelve authors are used as the authors of anonymised email texts in an authorship attribution experiment. Samples of various sizes are taken from the full sets of these twelve authors, and compared against their remaining emails and the full email sets of the other 175 Enron employees, in order to successfully attribute the email samples to the correct EEC12 author.

3.4.4 The eighty-author sample (EEC80)

Chapter 7 shifts the primary focus away from individual authors and towards groups of authors, particularly in identifying linguistic features which can discriminate between these different groups. To do this, an eighty-author sample of the Enron Email Corpus (hereafter EEC80) is used. EEC80 comprises 80 of the 176 employees in the overall corpus, 55,675 of the 63,369 emails in the corpus (87.86%) and 2,178,205 of the 2,462,151 tokens (88.47%) (Table 4). (A full breakdown of the EEC80 sample by author is given in Appendix 3).

Table 4. Breakdown of the EEC80 sample by occupation and sex

Occupation group	**authors**	**M**	**F**	**emails**	**tokens**
Presidents, CEOs and COOs	10	8	2	6,445	266,500
Vice Presidents	10	8	2	12,207	435,874
Managing Directors and Directors	10	4	6	8,466	407,047
Lawyers	10	4	6	10,945	407,646
Managers	10	9	1	6,184	197,518
Traders	10	6	4	5,718	207,321
Analysts, Specialists and Associates	10	7	3	4,903	225,365
Assistants	10	0	10	807	30,934
Totals	**80**	**46**	**34**	**55,675**	**2,178,205**

The EEC80 sample comprises 46 males and 34 females. Authors were divided into these two binary 'sex' groups on the basis of their names (all of which were straightforward to identify as being male or female). Males are represented by 32,129 emails and 1,181,409 tokens, compared with 23,546 emails and 996,796 for females. The sample is distributed across eight different occupation types in the corporation in decreasing hierarchical power, from Presidents, Chief Executive Officers (CEOs) and Chief Operating Officers (COOs) to administrative assistants, with ten authors in each occupation group. The higher number of males than females in EEC80 can be accounted for by the fact that some of the occupation groups in Enron (e.g. Presidents, CEOs and COOs, Vice Presidents, and Managers) comprise predominantly male employees. As in the compilation of the Pilot Corpus and the EEC12 sample, various sources were used to determine the official role of each individual participant in the corpus. First, Baucus and Grassley (2003) is an official document of the US Senate Committee on Finance, reporting the investigation of Enron regarding federal tax issues. Second, Creamer et al. (2009) is a paper by a research team from Columbia University, New York, investigating social hierarchy detection in the Enron email dataset, which gives information about the jobs of the individuals. Third, Priebe et al. (2010) is a description of a version of the Enron corpus designed by a research team at the Centre for Imaging Science at Johns Hopkins University, Baltimore, Maryland, as part of their scan statistics research on the corpus. Employees were only chosen for inclusion in EEC80 on the basis that at least two of these three sources agreed with regard to what their official job title or occupation was in Enron. The employees were included in this sample on the basis that they were the eighty authors with the largest datasets for their given job type. This EEC80 sample comprises no more than 80 authors because the aim was to have an equal number of authors across each of the eight groups, and for some occupations (e.g. Vice Presidents and assistants) there were no more than ten employees in the Enron Email Corpus. Nevertheless, although it includes fewer than half of the 176 authors, the EEC80 sample contains 88.47% of the total tokens in the Enron Email Corpus. As with any classification system, categorising individuals and job types by hierarchy was not unproblematic. First, as was noted with authors such as Arnold and Lavorato, an employee can be a trader (or lawyer), for example, as well as having a high status role in the company such as President or Vice President. In such cases, authors were categorised, here, on the basis of their highest stated role

in the three sources used. This resulted in authors such as Lavorato and Steffes, who were considered a trader and lawyer respectively in EEC12, being categorised as Vice Presidents in EEC80. The reason for this, beside the fact that EEC12 is concerned with linguistic individuals rather than groups, is that in Chapter 7, some comparison between occupation groups is to be made with regard to their seniority in the corporation. The ranking of job roles in the company was also complicated. While ranking some roles higher than others is straightforward, such as Presidents, CEOs, COOs, and Vice Presidents being highest ranked, and administrative assistants being lowest ranked, the boundaries in between others are less clear. For instance, some high ranking managers may, in the overall corporation, be superior to lower ranking lawyers. To overcome this, the hierarchy proposed here is informed by that of previous research into Enron (Rowe et al. 2007; Palus et al. 2010; Mitra and Gilbert 2012; Gilbert 2012) and is intended as an overall *ad hoc* ranking, and any unsubstantiated speculative generalisations with regard to influences on language use based upon it will be avoided in Chapter 7. As in EEC12, all of the emails from all of the sent folders were used for the 80 authors in this sample, and this results in some occupation groups such as Vice Presidents and lawyers having far more data than other groups, particularly assistants.

The EEC80 sample is comprised of 46 males and 34 females which, although not an exactly equal split, ensures that both sexes (male and female) are well represented within it. What is clear from Table 4 is that there is a strong relationship between sex and occupation in the EEC80 sample, particularly for certain roles. Of the top two most senior job categories, 16 of the 20 employees in the sample are men, as well as almost all of the managers, traders and analysts, associates and specialists. In contrast, there are more female than male directors/managing directors and lawyers, and all of the assistants are female.

The EEC80 sample compares favourably with other corpora used in authorship profiling research. On the one hand, it is far smaller than those studies that draw on existing reference corpora such as the *BNC* and the *International Corpus of Learner English* which include over one thousand authors and hundreds of millions of words (e.g. Koppel et al. 2005; Schler et al. 2005). On the other, it is similar in size to other email corpora used for profiling, such as Estival et al. (2007) which measures at around 10,000 emails and four million words, and the *Personae* corpus containing the writing of 145 authors and 200,000 tokens (Luyckx and

Daelemans 2008; Noecker et al. 2013). While not boasting the size of some corpora, the EEC80 sample shares with these smaller more specialised corpora the fact that it consists of individually identifiable authors. This allows the analyst to pinpoint particular authors who may skew results by using a particular linguistic feature either exceptionally rarely or frequently within a given social group. Also, because each author is represented by a substantial amount of data (between 605 and 148,518 tokens) across a number of texts (between 11 and 3,465 emails) an individual's linguistic behaviour can be profiled across emails, contexts and time. In turn, this corpus is suitable for both nomothetic (quantitative) and idiographic (qualitative, case study) approaches to profiling.

3.5 Chapter conclusion

The Enron Email Corpus is a unique dataset, providing an unparalleled insight into one of the largest and most innovative companies in US history. It is a dataset which offers a lot of opportunities for linguistics across many disciplines, including those researching corpus linguistics, pragmatics, workplace discourse and sociolinguistics. However, most of the research to date has been computational and mathematical in nature, and the dataset for the most part remains untouched by linguists. This research, along with Wright (2013) and Johnson and Wright (2014) is the first from the field of forensic linguistics to utilise this corpus, which holds many benefits for authorship analysis, given that it contains naturally-occurring language and sizeable amounts of data for individual identifiable authors. For the purposes of this research the corpus is used in three forms. The twelve-author EEC12 sample is used in the corpus stylistic analysis in Chapter 5 and the authorship attribution experiments in Chapter 6, and the EEC80 sample is used in the author profiling in Chapter 7. In addition, the full Enron Email Corpus is used as a reference corpus in Chapter 5 and a pool of 176 candidate authors in Chapter 6. The following section details the tools and techniques used in the analysis of the corpus and its various subsets throughout this study.

4 Methodology: tools and techniques

Methodology has a central role in this study, specifically the use of corpus linguistic approaches in combining qualitative and quantitative analysis of texts. This section introduces the corpus linguistic software—commercial and bespoke—used in the following chapters, as well as the simple statistical techniques applied and drawn upon throughout.

4.1 *Wordsmith Tools*

Wordsmith Tools version 5, developed by Mike Scott (Scott 2008a), is 'an integrated suite of programs for looking at how words behave in texts' (Scott 2008b). It is a member of the 'third generation' of corpus analysis tools (McEnery and Hardie 2012: 37), being introduced in the 1990s and 2000s, building on earlier tools such as *CLOC* (Reed 1978) and *Microconcord* (Scott and Johns 1993) used at the University of Birmingham and the *Oxford Concordance Program* (Hockey 1988). *Wordsmith* is available for purchase from Mike Scott's website (www.lexically.net) and is one of a number of commercially and freely available corpus linguistic software packages along with *AntConc* (Anthony 2014) and *Sketch Engine* (Kilgariff et al. 2014). *Wordsmith* was chosen for this study given the unparalleled features for the analysis of lexis and lexico-grammar required throughout this study, namely 'Keyword' analyses, concordance and cluster analysis and the 'Detailed Consistency' facility.

4.1.1 'Keyword' analysis

A 'keyword' analysis identifies those words 'whose frequency is unusually high in comparison with some norm' (Scott 2008b). To do this, the program compares two word lists (also created using *Wordsmith*), one for the corpus or text under study and a much larger one which serves as a reference list. The relative frequencies of words in these two lists are compared using the log-likelihood statistic (see Section 4.4.1) in order to identify those words which occur in the corpus under study more or less often than would be expected by chance in comparison with the reference corpus (Scott 2008b). In this study, a wordlist for the Enron Email Corpus is compared against a reference list of the 464 million word *Corpus of Contemporary American English* (*COCA*) (Davies 2012) to reveal those words which appear more frequently in the Enron Email Corpus than in American English generally. In addition, as in

previous research (Johnson and Wright 2014) a word list for the emails of one Enron employee can be compared with the rest of the Enron Email Corpus to find out which words they use more frequently than is expected in the Enron population from which they are taken.

Keywords 'provide a useful way to characterise a text or genre' and give a good indication of a text's or corpus's 'aboutness' (Scott 2008b). Keyword analysis is now a standard procedure for analysing and comparing texts in corpus linguistic approaches to a range of fields, including Critical Discourse Analysis (Baker et al. 2008 and 2013; Jeffries and Walker 2012; O'Halloran 2009), sociolinguistics and language variation (Barbieri 2008; Baker 2010; Murphy 2010), language teaching and learning (Scott and Tribble 2006; O'Keefe et al. 2007) and stylistics and literary studies (Archer et al. 2009; Culpeper 2009; Mahlberg and McIntyre 2011). For a recent survey on the applications of keyword analysis, see Archer (2009). Coulthard (1994: 32) effectively performed a keyword analysis when comparing the frequency of *then* in Derek Bentley's confession statement with that in the spoken subset of the *COBUILD* reference corpus. However, since then, only Hoover (2009) and Johnson and Wright (2014) pursue the use of a keyword approach to analysing authorial style. In an authorship context, relative word frequency has been the driving force behind the vast majority of stylometric research (See Chapter 2). A keyword analysis offers an alternative use of word frequencies; rather than relying on quantitative results alone, keywords offer a 'point of entry' for further linguistic analysis, as they do in other fields of corpus linguistic application:

> an examination of high-frequency words helps to indicate the main foci of a corpus in terms of indicating words or phrases the analyst might want to subject to further collocational and concordance analyses.
>
> (Baker 2010: 133)

4.1.2 Concordances, collocations and clusters

The 'Concord' feature in *Wordsmith* is used throughout this study to create concordance lines for specific search terms or phrases, which present all instances of that word or phrase used in context in the corpus or particular sub-corpus under analysis. An example of the output of a concordance result for the word *trade* in the Enron Email Corpus is given in Figure 4.

Figure 4. Screenshot of *Wordsmith* concord results for *trade* in EEC

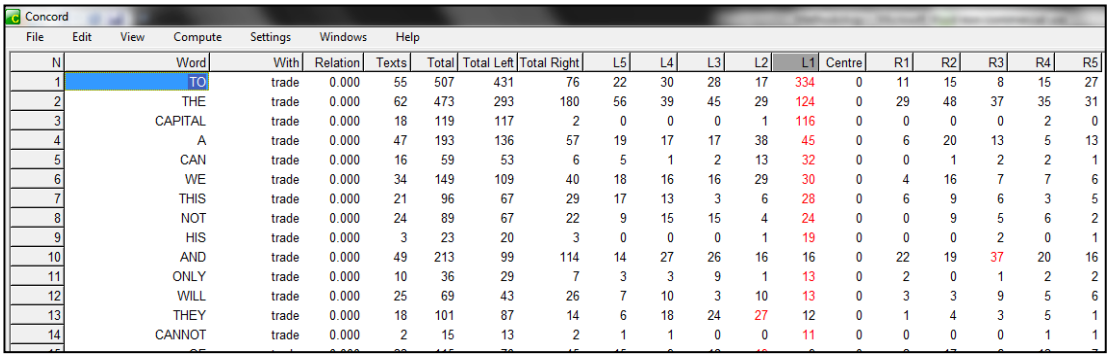
Concord

File Edit View Compute Settings Windows Help

N	Concordance	File
1	authorizing Mike G., Frank E., Keith H. and myself to trade west power. What do you think? Phillip <Message-ID: > I	allen-p-Dedup [Edited].txt
2	and Prebon). Please see the attached spreadsheet for a trade by trade list and a summary. We have also included a	allen-p-Dedup [Edited].txt
3	Prebon). Please see the attached spreadsheet for a trade by trade list and a summary. We have also included a summary	allen-p-Dedup [Edited].txt
4	on the West Desk. I do not need an ID that can only trade the West region. If possible, please modify simtrader7.	allen-p-Dedup [Edited].txt
5	to send? The message being received is that we should trade less aggressively and leave several million dollars of	allen-p-Dedup [Edited].txt
6	help? Phillip <Message-ID: > Stephanie, I need to be able to trade US Gas Spreads. Specifically the new west region	allen-p-Dedup [Edited].txt
7	shows why our transactional model of being one side of every trade is superior. #3. good liquidity first 3 years. okay liquidity	arnold-j-Dedup [Edited].txt
8	okay liquidity years 4-6. #4. calendar 2004-2008 maybe 1 trade a day. 70% chance Enron is one side. calendar	arnold-j-Dedup [Edited].txt
9	whom click on hub gas daily when they meant to trade nymex. Although I would like the nymex filter to bring	arnold-j-Dedup [Edited].txt
10	deserted and virtually all trading occurred electronically. I still trade on the exchange because we have credit issues with	arnold-j-Dedup [Edited].txt
11	don't follow closely, that I translated into a multimillion dollar trade for Enron. In return, I have agreed to have Enron	arnold-j-Dedup [Edited].txt
12	extremely rare as very few non-investment grade companies trade these types of products. Finally, on Friday Bill wanted to	arnold-j-Dedup [Edited].txt
13	these types of products. Finally, on Friday Bill wanted to do a trade that reduced his exposure to Enron. I gave Mike Maggi	arnold-j-Dedup [Edited].txt
14	exposure to Enron. I gave Mike Maggi the go ahead to do the trade without consulting credit. I do not believe that I acted out	arnold-j-Dedup [Edited].txt
15	credit. I do not believe that I acted out of line in approving this trade considering the circumstances. If you believe differently,	arnold-j-Dedup [Edited].txt
16	It's not your style to change views quickly as you tend to trade with a longer term view. I was out of line with the	arnold-j-Dedup [Edited].txt
17	placed, an error message occurs on the system saying error trade, price not available. 4. Is there a way to modify a limit	arnold-j-Dedup [Edited].txt
18	the amazing year you've had so far. Maybe you should come trade this... John <Message-ID: > I would like to participate in	arnold-j-Dedup [Edited].txt
19	has a long tail only on the positive p&l side. While such a trade in an efficient market has expected payout of 0, the	arnold-j-Dedup [Edited].txt
20	to EOL. It may speed up the process if they see they can trade pre-market. John <Message-ID: > I spoke to Vlady this	arnold-j-Dedup [Edited].txt

The Concord function can also compute and produce frequencies of all of the collocates of a given search word, ranging from the collocates five places before the word (L5) to five places after it (R5). Figure 5 shows the collocate results for *trade*, sorted by the frequency of L1 collocates, those which immediately precede *trade*. Using this output we find that *to trade* is the most frequent collocation in the EEC occurring 334 times in the corpus, followed by *the trade* and *capital trade*.

Figure 5. Collocate frequency for *trade* in EEC



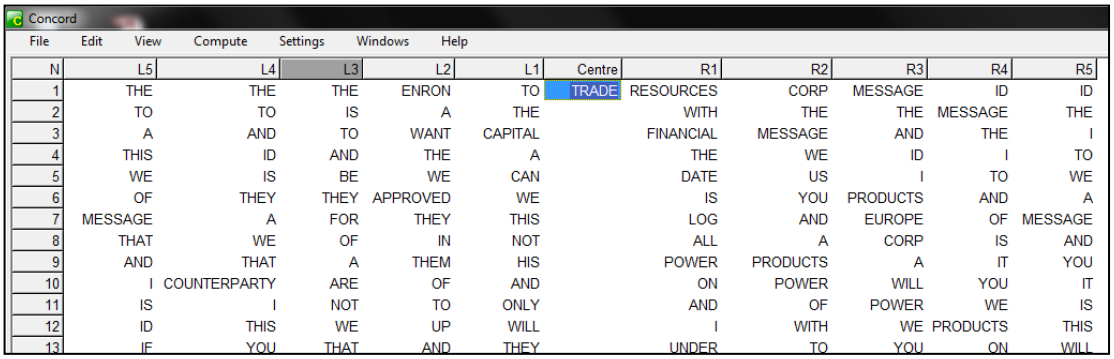
Concord

File Edit View Compute Settings Windows Help

N	Word	With	Relation	Texts	Total	Total Left	Total Right	L5	L4	L3	L2	L1	Centre	R1	R2	R3	R4	R5
1	TO	trade	0.000	55	507	431	76	22	30	28	17	334	0	11	15	8	15	27
2	THE	trade	0.000	62	473	293	180	56	39	45	29	124	0	29	48	37	35	31
3	CAPITAL	trade	0.000	18	119	117	2	0	0	0	1	116	0	0	0	0	2	0
4	A	trade	0.000	47	193	136	57	19	17	17	38	45	0	6	20	13	5	13
5	CAN	trade	0.000	16	59	53	6	5	1	2	13	32	0	0	1	2	2	1
6	WE	trade	0.000	34	149	109	40	18	16	16	29	30	0	4	16	7	7	6
7	THIS	trade	0.000	21	96	67	29	17	13	3	6	28	0	6	9	6	3	5
8	NOT	trade	0.000	24	89	67	22	9	15	15	4	24	0	0	9	5	6	2
9	HIS	trade	0.000	3	23	20	3	0	0	0	1	19	0	0	0	2	0	1
10	AND	trade	0.000	49	213	99	114	14	27	26	16	16	0	22	19	37	20	16
11	ONLY	trade	0.000	10	36	29	7	3	3	9	1	13	0	2	0	1	2	2
12	WILL	trade	0.000	25	69	43	26	7	10	3	10	13	0	3	3	9	5	6
13	THEY	trade	0.000	18	101	87	14	6	18	24	27	12	0	1	4	3	5	1
14	CANNOT	trade	0.000	2	15	13	2	1	1	0	0	11	0	0	0	0	1	1

To complement this, the Concord feature produces collocation 'patterns', which ranks L5 to R5 collocates in terms of how frequently they occur. For example, Figure 6 shows that the most common R1 collocates of *trade* are *resources, with, financial* and *the*, the most common R2 collocates are *corp, the,* and *we,* and so on. The appearance of *Message* and *ID* in these results indicates the end of messages with <Message ID>, marking the start of a new email in the corpus.

Figure 6. Collocation patterns of *trade* in EEC



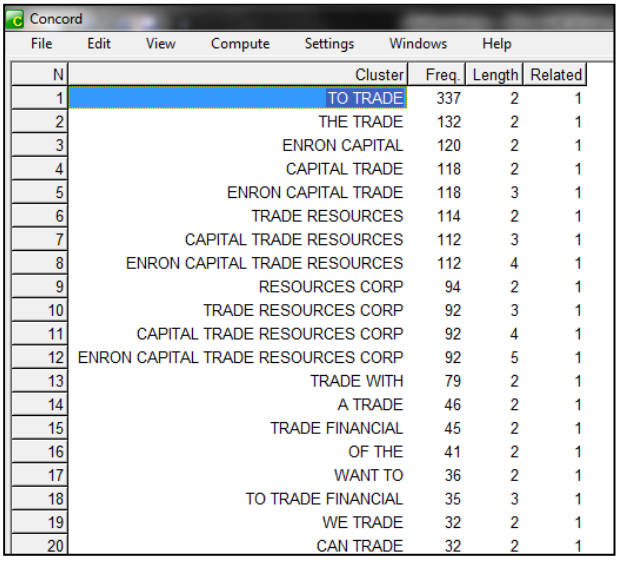
Concord

File Edit View Compute Settings Windows Help

N	L5	L4	L3	L2	L1	Centre	R1	R2	R3	R4	R5
1	THE	THE	THE	ENRON	TO	TRADE	RESOURCES	CORP	MESSAGE	ID	ID
2	TO	TO	IS	A	THE		WITH	THE	THE	MESSAGE	THE
3	A	AND	TO	WANT	CAPITAL		FINANCIAL	MESSAGE	AND	THE	I
4	THIS	ID	AND	THE	A		THE	WE	ID	I	TO
5	WE	IS	BE	WE	CAN		DATE	US	I	TO	WE
6	OF	THEY	THEY	APPROVED	WE		IS	YOU	PRODUCTS	AND	A
7	MESSAGE	A	FOR	THEY	THIS		LOG	AND	EUROPE	OF	MESSAGE
8	THAT	WE	OF	IN	NOT		ALL	A	CORP	IS	AND
9	AND	THAT	A	THEM	HIS		POWER	PRODUCTS	A	IT	YOU
10	I	COUNTERPARTY	ARE	OF	AND		ON	POWER	WILL	YOU	IT
11	IS	I	NOT	TO	ONLY		AND	OF	POWER	WE	IS
12	ID	THIS	WE	UP	WILL		I	WITH	WE	PRODUCTS	THIS
13	IF	YOU	THAT	AND	THEY		UNDER	TO	YOU	ON	WILL

The inclusion of these words in such results is useful for identifying whether a word is frequently used at the very start or very end of emails. Though they appear in these results here, because they are included in angled brackets they are excluded from all other *Wordsmith* results, such as keyword analysis and all *Jangle* results (Section 4.2 below). In addition to counting collocates and producing collocation patterns, the *Wordsmith* Concord function also produces a list of the most frequent word clusters found within the concordance results for a given node word, which may or may not include the node word itself. Settings can be adjusted to alter the length of clusters captured by the program. Figure 7, for example, shows the most frequent two to five word clusters found within the L5–R5 horizons of *trade*.

Figure 7. Most frequent two to five word clusters for *trade*



Concord

File Edit View Compute Settings Windows Help

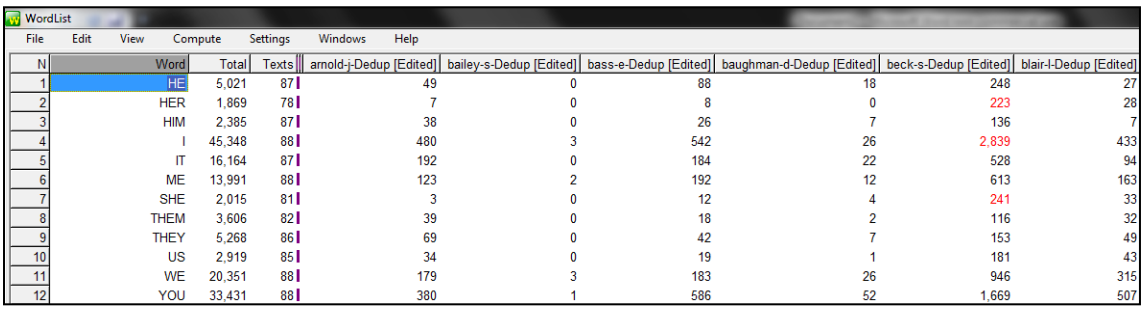
N	Cluster	Freq.	Length	Related
1	TO TRADE	337	2	1
2	THE TRADE	132	2	1
3	ENRON CAPITAL	120	2	1
4	CAPITAL TRADE	118	2	1
5	ENRON CAPITAL TRADE	118	3	1
6	TRADE RESOURCES	114	2	1
7	CAPITAL TRADE RESOURCES	112	3	1
8	ENRON CAPITAL TRADE RESOURCES	112	4	1
9	RESOURCES CORP	94	2	1
10	TRADE RESOURCES CORP	92	3	1
11	CAPITAL TRADE RESOURCES CORP	92	4	1
12	ENRON CAPITAL TRADE RESOURCES CORP	92	5	1
13	TRADE WITH	79	2	1
14	A TRADE	46	2	1
15	TRADE FINANCIAL	45	2	1
16	OF THE	41	2	1
17	WANT TO	36	2	1
18	TO TRADE FINANCIAL	35	3	1
19	WE TRADE	32	2	1
20	CAN TRADE	32	2	1

This collection of functions available in the Concord package of *Wordsmith* offers a range of ways to analyse how a particular word is used in a corpus, sub-corpus or by a particular author, in terms of the collocates it occurs with and the clusters it appears in. Throughout this study such analyses are performed and authors' collocational and clustering patterns and preferences are compared and contrasted both with each other and with the EEC generally.

4.1.3 Detailed consistency

'Detailed consistency' is a much less commonly employed feature within *Wordsmith* than Keyword and Concord. It enables the analyst to compare multiple word lists, and its main purpose is 'to help stylistic comparisons' (Scott 2008b) across them. This is a particularly useful function for this study, as it allows for word lists of multiple authors to be compared and contrasted with each other. Moreover, when using Detailed Consistency, a stop list of words which the analyst is interested in can be created and loaded, so that when the process runs it compares only the frequency of these words across the authors. This is the procedure used in the author profiling in Chapter 7 in counting the frequencies of various words and parts-of-speech. Figure 8, for example, shows the Detailed Consistency results for personal pronouns in the authors in the EEC80 sample. For this particular example, a list of all of the personal pronouns in the sample, identified as such by the CLAWS tagger (see Section 4.3), was loaded as a stop list and the frequency of these items was compared across all 80 authors. The results are presented in the Figure as produced by *Wordsmith*, with columns containing the authors and the rows containing the words. This procedure was followed for all of the parts-of-speech included in the author profiling analysis in Chapter 7.

Figure 8. Detailed Consistency results for personal pronouns in EEC80



WordList

File Edit View Compute Settings Windows Help

N	Word	Total	Texts	arnold-j-Dedup [Edited]	bailey-s-Dedup [Edited]	bass-e-Dedup [Edited]	baughman-d-Dedup [Edited]	beck-s-Dedup [Edited]	blair-l-Dedup [Edited]
1	HE	5,021	87	49	0	88	18	248	27
2	HER	1,869	78	7	0	8	0	223	28
3	HIM	2,385	87	38	0	26	7	136	7
4	I	45,348	88	480	3	542	26	2,839	433
5	IT	16,164	87	192	0	184	22	528	94
6	ME	13,991	88	123	2	192	12	613	163
7	SHE	2,015	81	3	0	12	4	241	33
8	THEM	3,606	82	39	0	18	2	116	32
9	THEY	5,268	86	69	0	42	7	153	49
10	US	2,919	85	34	0	19	1	181	43
11	WE	20,351	88	179	3	183	26	946	315
12	YOU	33,431	88	380	1	586	52	1,669	507

The three functions of *Wordsmith* described here are used throughout all chapters of this thesis, offering a variety of ways of analysing and comparing the lexico-grammar of the Enron Email Corpus and the authors who inhabit it. *Wordsmith* is used in combination with another computer program, which is described in the following section.

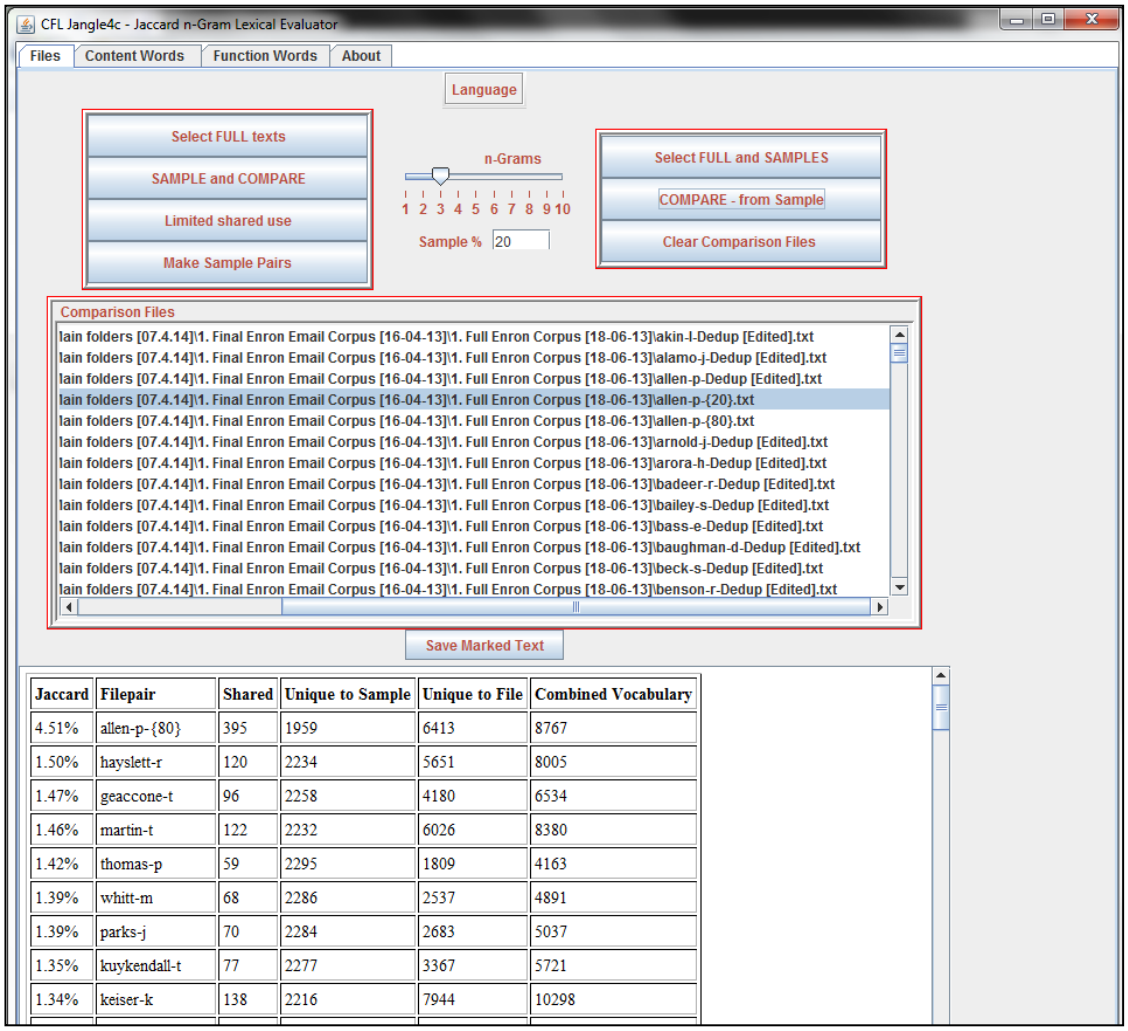
4.2 *Jaccard N-gram Lexical Evaluator* (*Jangle*)

The author identification experiments performed and reported in Chapter 6 of this thesis were run using a bespoke piece of Java-based linguistic analysis software developed by David Woolls of CFL Software Limited called *Jaccard N-Gram Lexical Evaluator* (Woolls 2013), referred to as *Jangle* from this point onwards.

The program is designed specifically for the purposes of this study, and is the product of developmental conversations with Woolls regarding the aims of this research. Figure 9 is a screenshot of its user interface. The program prepares for the authorship experiment by automatically generating random samples of emails for any one author, of any proportion the user requires, and separates these samples from the remainder of that author's emails. The screenshot shows, in the 'Comparison files' box, all of the plain text files containing the email sets for each of the authors in the corpus loaded into the program. As well as these files, there are two additional files for Allen: a random 20% sample of his emails (allen-p-{20}), and the remaining 80% (allen-p-{80}), which were created by using the 'Make Sample Pairs' option on the program's interface. In forensic casework terms, the 20% sample represents the anonymous 'questioned' or 'disputed' document(s), the 80% sample represents the 'known' writings of Allen (a suspect in this hypothetical case). The other 175 authors loaded into the program represent additional candidate authors of the 'questioned' text(s).

Once the random email sample has been created and extracted from the author under inspection (in this case Allen), the 'Compare – from Sample' option commands the program to run a series of pair-wise comparisons, with the sample file (allen-p-{20}) being systematically compared with all of the other files loaded into the program, including both the remaining emails of the author in question (allen-p-{80}) and the entire email sets of all the other Enron employees.

Figure 9. User interface of *Jangle* program and results for Allen



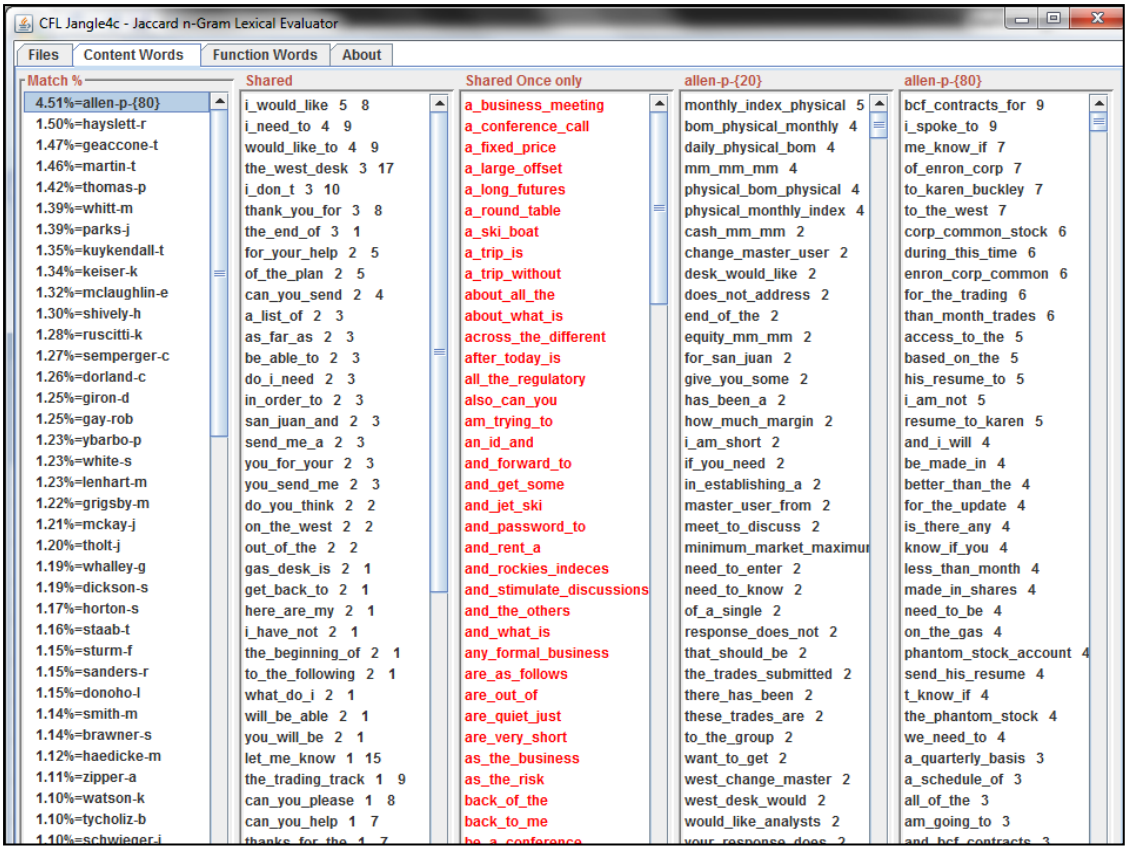
Jaccard	Filepair	Shared	Unique to Sample	Unique to File	Combined Vocabulary
4.51%	allen-p-{80}	395	1959	6413	8767
1.50%	hayslett-r	120	2234	5651	8005
1.47%	geaccone-t	96	2258	4180	6534
1.46%	martin-t	122	2232	6026	8380
1.42%	thomas-p	59	2295	1809	4163
1.39%	whitt-m	68	2286	2537	4891
1.39%	parks-j	70	2284	2683	5037
1.35%	kuykendall-t	77	2277	3367	5721
1.34%	keiser-k	138	2216	7944	10298

Each comparison measures the similarity between the sample set and the comparison set based on the number of word n-grams shared between them; in this case the length of n-gram chosen was trigrams (three words in length), but the program allows for much shorter and much longer n-grams to be used (n=1 to 10). The table in the bottom half of the user interface in Figure 9 shows the results of this example experiment. They show that the remaining sample of Allen's emails (allen-p-{80}) is more similar to the sample being tested than any of the email sets of any of the other 175 candidate authors. His remaining emails are 4.51% similar to the disputed sample. This low percentage is not surprising given that the sample being tested contains only 20% of Allen's emails. This is followed by a sharp decline to Rod Hayslett in second place with a similarity of 1.50%. In attribution terms, this result can be interpreted as a successful attribution, with Allen having been correctly identified as the author of the sample, as he achieved the highest Jaccard score. In

the same way as Juola (2013: 297), but unlike Grant (2013: 483), no attempt is made in this study to measure whether differences in Jaccard scores across pairs of texts are statistically significant or not. The only thing that is considered here with regard to successful attributions is that the correct author obtains the highest Jaccard score. There are methodological caveats to this approach (Section 6.5). However, to help avoid misleading results, the method is tested on a large number of different random email samples (Section 6.2). Jaccard's similarity co-efficient is explained in detail, along with the interpretation of these results, in Section 4.4.2 below. The Jaccard calculation is built into the program and is automatically performed, producing the result in the left-most column of the table in Figure 9.

As well as being able to automatically extract samples and run multiple Jaccard comparisons—across various n-gram lengths—in seconds, the main advantage of *Jangle* for linguistic analysis and authorship attribution is that it provides the analyst with the actual word n-grams operating behind the statistical results. Whereas with other stylometric programs (e.g. JGAAP by Patrick Juola) the files are processed and a set of statistics is returned, the Jaccard results offered by *Jangle* are supplemented with and supported by the linguistic evidence accounting for the statistics. Selecting the 'Content Words' tab on the program interface presents all of the word n-grams found in both the sample and the comparison text, as well as those found *only* in the sample and *only* in the comparison text (Figure 10). The results in the Figure are for the comparison of the sample allen-p-{20} with the comparison text allen-p-{80}, based on trigrams. They show that the trigram *I would like* is found shared between the sample and the comparison file multiple times. Although the Jaccard statistic is only concerned with the binary occurrence or non-occurrence of a feature in two sets (Section 4.4.3), *Jangle* shows us that this particular trigram occurs five times in the sample and eight times in the comparison file. Similarly, *I need to, would like to* and *the west desk* are all found multiple times in both sets of emails. At the same time, other n-grams such as *a business meeting, a conference call* and *a fixed price* are all also found in both the sample and comparison but only once in each. Meanwhile, the trigrams *monthly index physical, bom physical monthly* and *daily physical bom* are all found only in the sample, while *bcf contracts for, I spoke to* and *me know if* are all only found in the comparison set.

Figure 10. Word n-gram results provided by *Jangle*



Using these results we can explain the statistical results and identify which exact word n-grams are accounting for the Jaccard results produced in the first instance. *Jangle* only counts word strings within sentence boundaries, marked by a full stop, as opposed to the 'bag of words' approach used in other computational studies (Juola 2008: 253; Stamatatos 2009: 540) which consider word strings regardless of whether a full stop appears between them. The rationale for staying within sentence boundaries is that, this way, word n-grams identified contain semantic as well as lexico-grammatical information.

The *Jangle* procedure described here was repeated hundreds of times in the authorship experiments in this study with multiple authors, various sample sizes and different n-gram lengths. The n-gram results provided by the program are used to gain an insight into the characteristic and distinctive word n-grams of particular authors and to offer linguistic and stylistic explanations for the statistical results obtained.

4.3 Other tools

4.3.1 CLAWS part-of-speech tagger

The authorship profiling analysis in Chapter 7 required stop lists of all words belonging to particular word classes (e.g. prepositions, determiners, personal pronouns etc.) to be produced, and processed in the Detailed Consistency analysis using *Wordsmith* (see Section 4.1.3). In order to produce these stop lists, the ECC80 corpus was tagged for part-of-speech (POS) using Lancaster University's CLAWS tagger (Garside and Smith 1997). CLAWS is described as a 'hybrid' tagging system, combining probabilistic and rule-based elements in its tagging, choosing a 'preferred tag for a word by calculating the most likely tag in the context of the word and its immediate neighbours' (Garside and Smith 1997; Rayson and Garside 1998). The CLAWS tagger was preferred over other available POS taggers (e.g. the Stanford NLP tagger, Toutanova et al. 2003) for a number of reasons. CLAWS is a freely available web-based system in which texts are inputted for tagging using a web browser (e.g. Internet Explorer, Safari, Google Chrome), and the tagged text is produced and returned to the user instantly, without any additional training of the tagging algorithm. Trained on approximately 100 million words in the *BNC*, the CLAWS system consistently operates with an accuracy rate of 96–97% in terms of correctly tagging lexical items and a mis-tagging error rate of only 1.5% (Leech et al. 1994: 625; Garside and Smith 1997: 119; see also the CLAWS website at ucrel.lancs.ac.uk/claws/). The tagger is monitored and maintained regularly as part of Rayson's (2003) wMatrix corpus linguistic software tool, and is widely used in current linguistic research (e.g. Leech et al. 2012; Brysbaert et al. 2012; Potts and Baker 2012; Davies and Gardner 2010).

Once the EEC80 corpus had been tagged using CLAWS, the results were exported into an Excel workbook, and the tagging results were manually examined for erroneous tags. When the errors had been removed or corrected, the results for twenty parts-of-speech (Table 5) were extracted. These parts of speech are included in the author profiling analysis as they include function words (e.g. articles, determiners, prepositions) as well as content words (e.g. nouns, words, adjectives). From these results, stop lists were created for each of the twenty parts-of-speech, and processed using the Detailed Consistency tool in *Wordsmith*, to count and

compare the frequencies of these words in the data for authors in EEC80 (Chapter 7).

Table 5. Parts-of-speech tagged and included in authorship profiling analysis

CLAWS tag*	POS	Example
AT0	articles	*the, a, an*
DPS, DT0	determiners	*these, some, your, their*
PNP	personal pronouns	*you, them, ours*
PNI	indefinite pronouns	*none, everything*
PNX	reflexive pronouns	*itself, ourselves*
ITJ	interjections	*oh, yes, hmm*
CJC, CJS	conjunctions	*and, or, although, when*
VM0	modal verbs	*can, would, will*
PRP, PRF	prepositions	*for, above, to, of*
AVQ, DTQ, PNQ,	*wh-* words	*who, how, why, when*
AJ0	adjectives	*good, old*
AJC	comparative adjectives	*better, older*
AJS	superlative adjectives	*best, oldest*
AV0	adverbs	*often, well, longer, furthest*
NN1	singular nouns	*pencil, goose*
NN2	plural nouns	*pencils, geese*
VVB, VVI	base form verbs	*take, live*
VVD	past tense verbs	*took, lived*
VVG	*-ing* verbs	*taking, living*
VVZ	*-s* verbs	*takes, lives*

*Tag list and POS codes key here: http://ucrel.lancs.ac.uk/claws5tags.html

4.3.2 *SPSS*

Statistical Package for the Social Sciences (SPSS, version 21) (IBM Corp 2012) is 'comprehensive, easy to use set of data and predictive analytics tools for business users, analysts and statistical programmers'.

SPSS can be used by linguists to analyse the relationship or non-relationship between variables, either linguistic or social. The software allows the analyst to perform a wide range of statistical tests and calculations, both descriptive and

inferential, and produces tabulated and graphical results. There is a precedent for using *SPSS* in forensic authorship analysis research in both data manipulation and statistical testing. Grant (2004; 2007) used *SPSS* to randomly select texts in his authorship attribution experiments, to perform descriptive statistics (means, standard deviations) and to perform discriminant function analysis. *SPSS* has also been used to perform discriminant function analysis by Turell (2010), Queralt and Turell (2012) and Chaski (2013). *SPSS* has not only been used to run discriminant analysis, though. Tomblin (2013: 82) uses the package to run a number of statistical tests: *t*-tests, Mann-Whitney U tests, Kruskal-Wallis tests, log-linear analysis and Kolmogorov-Smirnov tests.

In this study, *SPSS 21* is used in Chapter 7 to calculate statistics which compare mean frequencies of variants and variables across authors—Mann-Whitney U and Kruskal-Wallis—explained in Section 4.4.2.

4.4 Statistics

The title of this thesis grandly pits statistics against stylistics, yet the statistical techniques used within this study are relatively straightforward ones. This is a deliberate decision. No attempts are made to replicate or apply more complex statistical procedures used in other authorship attribution research, such as the widely-employed discriminant function analysis (Grant 2007; Turell 2010; Rico-Sulayes 2011; Chaski 2013), multivariate cluster analysis (Hoover 2003; Burrows 2005; Labbé 2007), principal component analysis (Baayen 1996; Binongo and Smith 1999; Grant and Baker 2001; Savoy 2012) and regression analysis (Koppel et al. 2011; Argamon and Koppel 2013; Rashid et al. 2013).

The reason that these more complex statistical techniques have not been employed here—besides the risk of a non-expert misusing them—is that there is currently a great amount of research effort, both in forensic authorship analysis and computational stylometry, towards testing and evaluating these various approaches. The position of this thesis is that a greater contribution to the field is to test the power and utility of more straightforward statistical techniques. This is fuelled by the desire to combine statistical and linguistic evidence in such a way that is accessible and understandable by lay judges and juries. In outlining the short-term aspirations for forensic authorship attribution Cheng (2013: 547) argues that the

testimony of the forensic linguistic expert 'must enlighten more than distort or confuse'. He continues:

> Statistical methods always have underlying assumptions and potential problems, and asking jurors (or even opposing counsel) to ferret out the distortions created by flawed models is unrealistic. Unless the method is so well-trodden and well-accepted that a jury can essentially use its results uncritically, I worry that statistical models in this context may distort more than illuminate.

The aim here, therefore, is to use more simple statistical procedures, that allow evidence to be quantifiable, and results to be robust and empirical, but that are straightforward enough to be understood and relied upon by legal decision-makers. The statistical techniques used in this study are described here.

4.4.1 Log-likelihood

In a forensic linguistic context, the use of the likelihood statistics has been most widely applied as likelihood ratios in forensic phonetics and voice comparison (Rose 2006; 2013). Specifically, they have been used in relation to the relevance of forensic evidence, and this has recently been explored in the context of authorship attribution (Ishihara 2014). The forensic linguistic applications of likelihood ratios have been borrowed from forensic scientific fields more broadly (e.g. Evett 1993). For example, Lucy (2005: 120) discusses the use of likelihood ratio calculations in a case where it was uncertain whether a piece of evidence recovered from the scene of a crime had been left by the offender or not. Put simply, Lucy (2005: 133) describes a likelihood ratio as 'an unambiguous and easily interpretable quantity which expresses the persuasive power of evidence'. He goes on to say that likelihood ratios can be verbalised in statements such as 'it is 265 times more likely that the evidence would be observed were the suspect the perpetrator than were some other person the perpetrator' (Lucy 2005: 133). It is here where the intersection between likelihood ratios in forensic fields and likelihood ratios in authorship attribution exists. It is in order to make such statements about the power of evidence that Coulthard (2010: 482) makes use of likelihood ratios. By comparing the frequency of linguistic variants in a questioned text message in a forensic case with the frequency of those variants in a 'representative sample' of text messages from a relevant population of writers, the analyst can calculate the likelihood that a text would be in a particular form if the writer in question had or had not written it. Such a procedure, it is

argued, would allow the forensic linguist to make statements akin to Lucy's (2005: 133) above, such as:

> a consideration of all the linguistic evidence in this single text message shows that it is some 26 times more likely that the text message would be in this form if the accused had sent it than if she had not.
>
> (Coulthard 2010: 483).

Wright (2013) also uses this kind of likelihood measure to calculate the likelihood that a particular Enron email would contain a certain greeting or farewell based on the rarity or frequency of these variants in the Enron population. It was found that some greetings are between 200 and 500 times more likely to be found in an email written by a particular author than by any other employees in the corpus.

The use of likelihood, specifically *log*-likelihood, in this thesis is related to but different from that already being used in forensic linguistics. Log-likelihood is a well-established statistical technique employed in corpus linguistics. Log-likelihood is the statistic used by *Wordsmith* to calculate keywords, the process for which is described in Section 4.1.1 above. The use of the log-likelihood measure to identify key words hinges on the comparison of corpora, and the relative frequencies of words across these corpora. The use of the log-likelihood statistic to identify salient lexical items in a text or corpus was first proposed by Dunning (1993) and later developed and evaluated by Rayson and Garside (2000), Rayson (2003) and Rayson et al. (2004). For any word in any two frequency lists, the log-likelihood statistic is calculated by the software being used constructing a contingency table as in Table 6.

Table 6. Contingency table for log-likelihood calculation
('-' = minus, '+' = add)

	Corpus 1	**Corpus 2**	**Total**
Frequency of word	a	b	a+b
Frequency of word not occurring	c-a	d-b	c+d-a-b
TOTAL	c	d	c+d

In this table, the value 'c' corresponds to the total number of words in corpus one, and 'd' corresponds to the number of words in corpus two (N values in the formula below). The values 'a' and 'b' are called the observed values (O) of the word in question. The 'expected' values (E) of the word are then calculated according to the following formula:

$$E_i = \frac{N_i \sum_i O_i}{\sum_i N_i}$$

In relation to the contingency table N1 is 'c', and N2 is 'd'. So, for this hypothetical word, E1 = c × (a+b) / (c+d) and E2 = d × (a+b) / (c+d). E1 and E2 are then fed into this formula which calculates the log-likelihood value:

$$-2\ln \lambda = 2\sum_i O_i \ln\left(\frac{O_i}{E_i}\right)$$

This equates to calculating log-likelihood as = 2×((a×ln (a/E1)) + (b×ln (b/E2))), with 'ln' referring to the 'natural log' of the number. This description of the log-likelihood calculation is adapted from Rayson (2003: 96–97). Put simply, the log-likelihood calculation compares the relative frequency of a particular word in the corpus under analysis and compares that with its relative frequency in the reference corpus. Log-likelihood features only in Chapter 5 of this study. First, it is used implicitly in the keyword analysis using *Wordsmith.* Based on the above formula a word will appear in a keyword list if it is unusually frequent in the corpus under analysis, by comparison with what one would expect on the basis of the larger word-list (Scott 2008b). Following that, log-likelihood is used in the analysis of *I* in Chapter 5, in the evaluation of 'Base Rate Knowledge'. It is used to measure the difference in the frequency with which an author uses *I* and its collocates when compared with how frequently it and its collocates appear the in the Enron corpus generally. The higher the log-likelihood value, the more significant is the difference between the two frequencies. The log-likelihood value obtained can be compared against a set of 'critical values' in such a way that we can measure whether an author uses a particular word or collocate statistically significantly more frequently than is expected in the Enron population. The critical values are as follows:

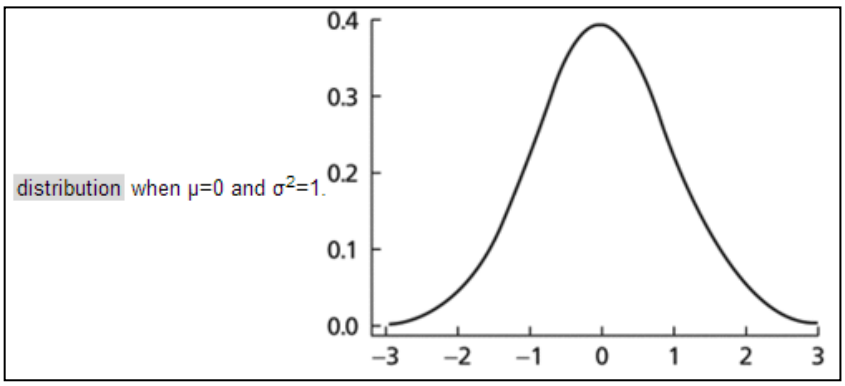
> 95th percentile; 5% level; $p < 0.05$; critical value = 3.84
> 99th percentile; 1% level; $p < 0.01$; critical value = 6.63
> 99.9th percentile; 0.1% level; $p < 0.001$; critical value = 10.83
> 99.99th percentile; 0.01% level; $p < 0.0001$; critical value = 15.13
>
> (Rayson et al. 2004: 7)

The null hypothesis in any test is that there is no difference between how frequently a particular word is used by the individual author in question and how frequently it appears in the corpus in general. If the log-likelihood value produced for any one word being compared across corpora is higher than 3.84 then this null hypothesis can be rejected. In Chapter 5, the analysis is concerned with focusing on those words and collocates that are 'distinctive' of individual authors. In order to make a claim for distinctiveness, then, the more significant the difference between an author's frequency and the overall frequency in the corpus, the better. For this reason only those words for which the log-likelihood value is higher than 15.13 are considered. At this level ($p<0.0001$), the difference between frequencies can be considered highly statistically significant, as we can be 99.99% confident that the difference observed between the individual author and the Enron Corpus is a true difference, and not due to chance. The reason that the significance level for log-likelihood has been set so high, as opposed to the more popular $p<0.05$ level (see Section 4.4.2), is because with this stricter threshold we can be more sure that the frequency with which an author uses a particular feature is truly distinctive of their writing style. These log-likelihood calculations for *I* collocates conducted outside *Wordsmith* are performed using the log-likelihood calculator spreadsheet made available for download by Paul Rayson at ucrel.lancs.ac.uk/llwizard.

4.4.2 Mann-Whitney U and Kruskal-Wallis

Mann-Whitney U and Kruskal-Wallis tests are used in Chapter 7 of this study. In statistics, a variable that has a 'normal' distribution will be represented by a symmetrical bell-shaped curve on a histogram, in which the mean and median are equal and are at the peak of the curve, with 95% of its values lying within two standard deviations of the mean (Figure 11).

Figure 11. Symmetric bell-shaped curve for normally distributed data. μ = mean, σ = standard deviation. (Source: Clapham and Nicholson [2009])



When data is normally distributed, the means of different groups within the population for a particular variable can be compared using 'parametric' statistical tests, such as *t*-tests when comparing two groups (e.g. male/female) or analysis of variance tests (ANOVA) for comparing more than two groups (e.g. different age groups). However, unlike height or weight in a population, linguistic variables are rarely normally distributed in this way (Barnbrook et al. 2013: 89; Phakiti 2010: 45) and this is the case of the linguistic variables in this study. Figure 12 and 13 show the distributions of *please* and *him* in the EEC80 sample.

Figure 12. Histogram and distribution of *please* in EEC80

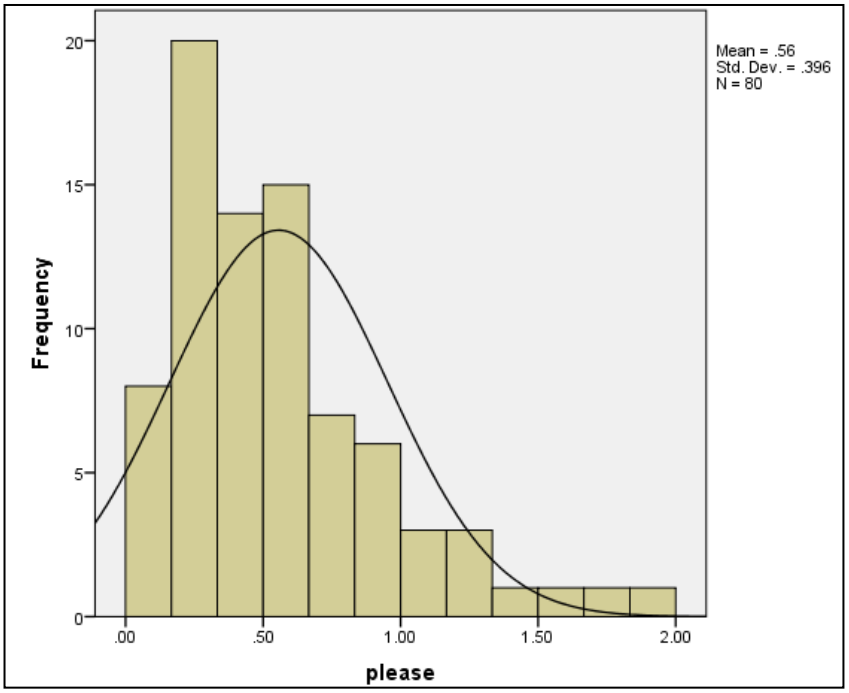
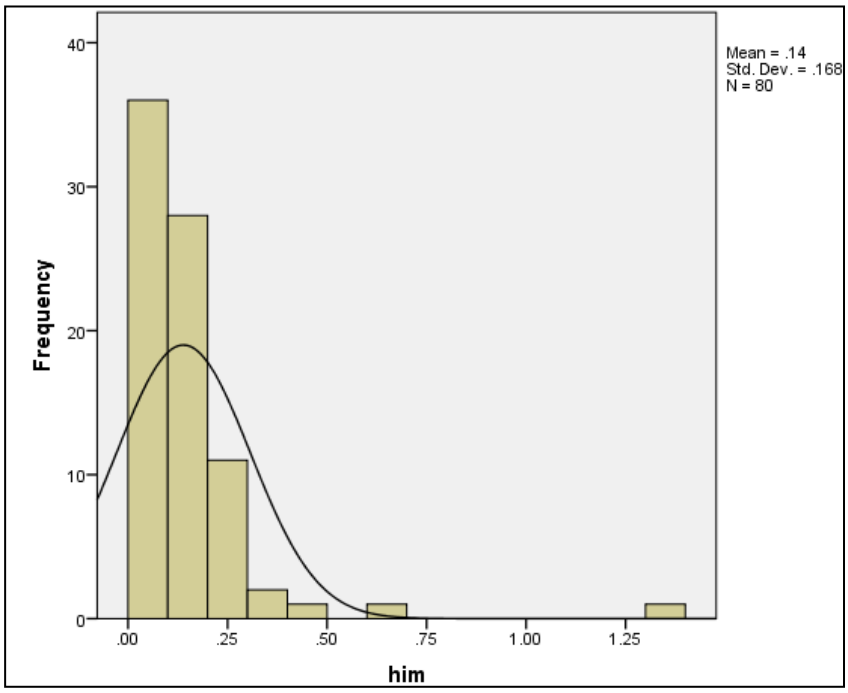


Figure 13. Histogram and distribution of *him* in EEC80



The distributions of both of these words are 'positively skewed' as there are few high values, with the majority of the values falling towards the lower side of the scale, making the mean higher than the median and producing an asymmetrically long tail to the right of the chart. The implication of this non-normal distribution of linguistic variables is that in the author profiling analysis in Chapter 7, normal 'parametric' tests (*t*-tests and ANOVA) cannot be used, as they assume normal distribution. Instead, non-parametric alternatives Mann-Whitney U and Kruskal-Wallis are used.

The Mann-Whitney U test is the non-parametric alternative of a *t*-test when comparing the means of two groups, in this case, male authors and female authors. Rather than using actual values, the Mann-Whitney test ranks the data for each variable. In this study, the test ranks individual authors based on how frequently they use a particular linguistic variable, paying no attention to whether the author is male or female. The lowest scoring author is ranked with a 1, the next lowest with 2 and so on. If there is a difference between male and female usage, then most of the high ranks will belong to one group, and most of the low ranks to the other. If there is no difference between the groups, then the ranks will be more evenly distributed. The Mann-Whitney 'U' statistic reflects the difference between the rank totals for the two groups using the following formula:

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

In this equation, n_1 and n_2 are the number of people in group 1 (males) and group 2 (females), and R_1 is the rank total for the most high ranking group (either males or females). The smaller the 'U' value produced the less likely it is that the difference between the two groups has occurred by chance. So, unlike with other statistics, the lower the Mann-Whitney value, the better for identifying statistical difference between groups. The null hypothesis in the author profiling analysis is that there is no difference between male and female Enron authors in their use of a particular linguistic feature. However, if the Mann-Whitney U value produced for any given linguistic variable is significant at the $p<0.05$ level, then the null hypothesis is rejected. In other words, a statistically significant difference between males and females has been found. The significance level of $p<0.05$ is conventional in statistics (see Grant and Baker 2001: 75 for a discussion of this in a forensic context). However, it is less strict than the $p<0.0001$ level applied with the log-likelihood statistic (see Section 4.4.1 above). The reason for this is so that the analysis can tease out as many significant differences between male and female authors as possible, rather than being as strict in identifying author-distinctive use in the analysis of *I* in Chapter 5. Mann-Whitney U is preferred over log-likelihood in the author profiling analysis. Recent linguistic research (Brezina and Meyerhoff 2014) has identified that, because Mann-Whitney takes into account *within*-group differences (as well as *between*-group ones), it produces more meaningful and accurate results than log-likelihood when comparing linguistic results across social groups. Such within-group differences are less important when using log-likelihood to compare an individual author against the rest of the corpus as outlined above (Section 4.4.1). Furthermore, Mann-Whitney tests have been used in previous author attribution (e.g. Kredens 2002; Grant 2013; Larner 2014) and author profiling (e.g. Argamon et al. 2003) research. The formula and description of the Mann-Whitney U statistic here are for reference only, as the tests themselves are performed automatically using *SPSS*.

In the same way as a *t*-test assumes normal distribution of data and cannot be used to compare males and females in the EEC80 sample, so too does ANOVA, and so cannot be used to compare more than two groups in the EEC80 sample data.

Therefore, the non-parametric alternative Kruskal-Wallis method is used alongside Mann-Whitney U in Chapter 7, to compare the frequency with which authors with different occupational roles in Enron use linguistic variables. Whereas Mann-Whitney U compares the means of just two sexes, Kruskal-Wallis compares the means across multiple occupational groups. In the same way as with Mann-Whitney, Kruskal-Wallis begins by ranking the frequencies of a given variable for all individuals, ignoring which occupation group they belong to; the author with the lowest relative frequency will get the lowest rank (1) and so on. The ranks are then totalled for each of the groups (in this case eight) and the Kruskal-Wallis value *'H'* is found using the following formula:

$$H = \left[\frac{12}{N(N+1)} * \Sigma \frac{Tc^2}{n_c} \right] - 3 * (N+1)$$

In this equation *N* is the total number of participants (all groups combined), in this case 80, *Tc* is the rank totals for each group, and *Nc* is the number of people in each group, in this case 10. If the '*H*' value is greater than 14.07, which is the critical value relevant for tests with eight groups, then the difference between the groups in their use of the particular linguistic feature in question is statistically significant at the level $p<0.05$. As with Mann-Whitney, all Kruskal-Wallis calculations are performed by *SPSS*. If such a result is obtained, it tells us that one of the groups' usage of the linguistic feature differs significantly from at least one of the other groups'. What it does not tell us is *which* groups in particular differ. In order to find this out, the total ranks for each of the groups are checked, to identify which of the eight occupation groups uses the feature in question the most frequently. In the analysis in Chapter 7, the feature is considered to be distinctive, or at least characteristic, of the group with the highest total rank. The Kruskal-Wallis statistic has been used in authorship research (e.g. Larner 2014), but less commonly than ANOVA or even Mann-Whitney. However, it has been more widely used in forensic case linkage (e.g. Woodhams and Toye 2007; Tonkin et al. 2008) and in corpus linguistics (e.g. Ho-Abdullah 2010; Titak and Roberson 2013).

4.4.3 Jaccard's similarity coefficient

In the authorship attribution experiments performed and reported in Chapter 6, the similarity between extracted email samples and comparison texts is measured using Jaccard's similarity coefficient (also known as 'Jaccard index' or 'intersection distance'). It is a simple calculation which measures similarity between any two datasets *A* and *B* by considering the fraction of the data that is shared between the two datasets (A∩B) as a proportion of all the data available in the union of the two sets (A∪B) (Naumann and Herschel 2010: 24), using the formula:

$$J(A,B) = \frac{A \cap B}{A \cup B}$$

Jaccard's coefficient has its origins in ecology (Jaccard 1912) and the related fields of marine biology and agricultural sciences, in which it is used to compare environmental sites based on number of shared flora and fauna species and functional traits (Izak and Price 2001; Bremner et al. 2003; Pottier et al. 2013; Tang et al. 2013). As an extension of this it has been used as a similarity measure in forensic psychology and crime analysis as part of the comparison of sites and offences in the behavioural linking of crimes and cases (Bennell and Jones 2005; Woodhams et al. 2007 and 2008; Markson et al. 2010). More recently, the use of Jaccard has gained momentum in data mining and document comparison research (Rajaraman and Ullman 2011; Deng et al. 2012; Manasse 2013) in the task of finding textually similar documents in large datasets. Jaccard was first introduced into forensic authorship analysis by Grant (2010: 518) as he applies it to the analysis of text messages in the Jenny Nicholl murder case. Since then it has been used by Wright (2012) in the attribution of Enron emails (pilot study, see Section 4.5), Juola (2013) in the attribution of newspaper article in an asylum case in an immigration court, and Larner (2014) in the evaluation of formulaic sequences as markers of authorship in the attribution of anonymous personal narratives.

The experiment in Chapter 6 aims to systematically and rigorously test the usefulness of Jaccard in measuring the similarity between email samples of various sizes and with over a hundred candidate authors, as well as with different word n-gram lengths. As explained in Section 4.1, each attribution experiment in Chapter 6 runs a series of pair-wise comparisons in which the sample in question is compared

against either the remaining emails of the author from which the sample is taken, or the email set of another Enron author. In terms of the Jaccard formula, dataset *A* is the questioned sample and dataset *B* is the comparison set. In each test, the similarity between *A* and *B* is based on the number of items—word n-grams—found in both sets, divided by the number of total number of items in the two sets combined. For the purposes of this analysis, the Jaccard formula can be represented as:

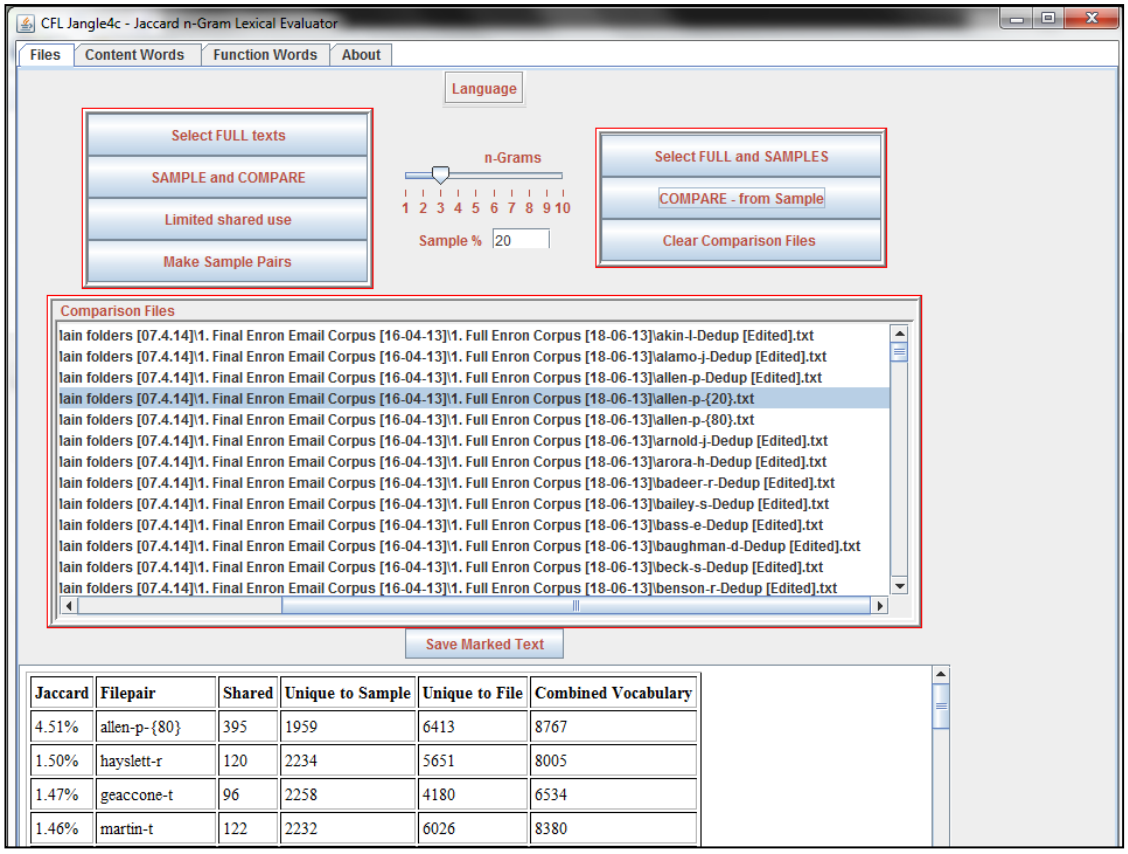
$$\frac{shared\ items}{shared\ items + items\ unique\ to\ sample + items\ unique\ to\ comparison\ file} \times 100$$

Jaccard is a binary correlation analysis in that it hinges on the appearance or non-appearance of a particular word n-gram in the two samples compared, rather than how frequently it occurs. Jaccard normally produces results between zero and 1, with zero indicating complete dissimilarity and 1 indicating that the two datasets are identical (Grant 2010: 518). However, in the interests of clarity, the results in this study have been multiplied by 100 and are expressed as percentages, so that 0% indicates that any two sets are completely different and 100% indicates that the datasets are identical. The example of attributing Allen's 20% sample used in the explanation of *Jangle* above (Section 4.1) can be revisited here as an explanation of the Jaccard statistic. The *Jangle* interface and results are reproduced in Figure 14 for reference. The numbers in the columns of the results produced by *Jangle* are used in the Jaccard statistic, for example in the comparison of the sample (allen-p-{20}) with Allen's remaining emails (allen-p-{80}):

i. Shared items (A∩B) = **395**
ii. Combined items (A∪B) = 395 + 1,959 + 6,413 = **8,767**
iii. Jaccard = **0.0451**
iv. x 100 = **4.51%**

There are 395 different trigrams that are found in both the sample and the comparison set, from a total of 8,767 different trigrams found in the two sets combined. With a possible range of between 0% and 100% Allen's remaining emails are similar to the sample to a degree of 4.51%. The important point here is that this Jaccard score is higher than any of the other pair-wise comparisons; Rod Hayslett was scored as being second most similar to Allen's sample, achieving a Jaccard similarity score of only 1.5%. In this example test, because the comparison

Figure 14. User interface of *Jangle* program and results for Allen



Jaccard	Filepair	Shared	Unique to Sample	Unique to File	Combined Vocabulary
4.51%	allen-p-{80}	395	1959	6413	8767
1.50%	hayslett-r	120	2234	5651	8005
1.47%	geaccone-t	96	2258	4180	6534
1.46%	martin-t	122	2232	6026	8380

file that was measured as being most similar to the sample was the remaining emails of the author from whom the sample was taken (Allen), this is a successful attribution. Again, because the Jaccard statistic is built into the *Jangle* program, hundreds of comparisons and Jaccard scores are computed in a matter of seconds.

There are a number of advantages in using Jaccard for the methodology proposed here. One of these advantages is that the occurrences of two absence scores or joint non-occurrences have no effect on the overall similarity rating (Woodhams et al. 2007: 18; Grant 2010: 518). That means that in the comparison of two authors' writing, the fact that they both *do not* use a particular word n-gram does not contribute to their level of similarity. This is necessary in an analysis such as this, as it is not surprising (in fact it is absolutely certain) that some of the same word n-grams will be absent from both of their writing, and so inclusion of such a factor would drive their similarity scores artificially high. Furthermore, given the simple nature of this calculation, and the ability for it to be applied quickly to any number of pair-wise author comparisons, it makes an attractive statistical tool for forensic linguists. Another advantage is that, because the shared data is considered

in proportion to the union of the data in the Jaccard formula, the measure is not dependent or sensitive to dataset size, so two authors represented by varying sizes of email data can still be compared accurately. This will be particularly important in forensic casework in which there may be substantial size discrepancies between known and disputed documents. Finally, the use of Jaccard as part of *Jangle* ensures that the actual word n-grams items operating behind, and accounting for, the Jaccard scores can be identified and explored to qualitatively examine any emerging patterns in authors' lexico-grammatical choices, and thus offer some linguistic explanation for the statistical results obtained.

Woodhams et al. (2007) and Macleod and Grant (2012: 214) offer a 'more robust' extension of Jaccard, called 'Delta-S' (not to be confused with Burrows' [2002] 'Delta'). Delta-S can usefully weigh the variables within a Jaccard calculation as being related to one another, which allows it to recognise 'similar but not identical stylistic choices to be represented in the final similarity metric' (Macleod and Grant 2012: 213). Although this is an attractive prospect, the use of Delta-S requires a taxonomy or hierarchy of variables for consideration, comprising multiple levels of increasing specificity. For example, Macleod and Grant's (2012: 218) feature categorisation system includes, within the 'Message' level, a 'Lexis' level, and within the 'Lexis' level, an 'Initialism' level, and then within that there are alternative options of 'syllable', 'single', 'compound' and 'phrase'. Although this may work for short form messages such as tweets and the consideration of many levels of variables, such a taxonomic or multi-level approach to lexico-grammatical similarity would necessarily include some kind of *ad hoc* semantic coding system. This would immediately place too many human restrictions and manipulations on the data, which would inevitably influence results. As such, the Delta-S metric could not be as effective or reliable as Jaccard for the purposes of this study. Indeed, Grant (2013) has since preferred Jaccard over Delta-S. Furthermore, Woodhams et al. (2007) considered only 16 offences in their investigation of case linkage. Similarly, Macleod and Grant (2012: 219) coded a maximum of 100 tweets per author. Although 100 is a fairly large amount of tweets, it is manageable qualitatively. In contrast, if a taxonomic approach was taken to lexico-grammar, it would be far more difficult to qualitatively code the tens of thousands of items in the Enron Email Corpus. Thus, overall, Jaccard is considered a more suitable alternative statistical

measure for calculating the level of similarity in the lexico-grammatical patterns exhibited in questioned email samples and known comparison sets.

4.5 Attribution experiment pilot study

Before the full attribution experiment reported in Chapter 6 was performed, a pilot was undertaken using the Pilot Corpus discussed above (Section 3.4.2). Running a pilot study offers a number of methodological advantages. Primarily, piloting is invaluable in 'assessing the practicality of the main study' (Rungruangthum et al. 2011: 32), serving as a 'dress rehearsal' for the full data collection procedures (Dörnyei 2007: 75) and uncovering any potential problems in the main study (Mackey and Gass 2005: 43). For these reasons, in preparation for the authorship attribution experiments in Chapter 6, a pilot authorship attribution experiment was run. This entailed collecting and preparing a sample of four authors (see Section 3.4.2) and ten different random sets of 100 emails were extracted from each of these authors. Using Jaccard and individual words (unigrams), these extracted sets were compared against the remaining emails of the author in question and the full email sets of the other three authors. The pilot experiment was very successful; the correct author of the 100 email sample was successfully identified in 39 of the 40 tests, a success rate of 97.5% (Wright 2012). That is, the correct author obtained the highest Jaccard similarity score in 39/40 tests.

The results of this study indicated that measuring lexical similarity between samples and comparison sets, taking into account whole vocabularies rather than just function words (as is most common in authorship research), held potential for authorship attribution. They also provided evidence to suggest that Jaccard's coefficient was a useful statistical tool for measuring such similarity. The pilot study was also helpful in determining the size of the samples to be extracted and attributed in the main study. For the four traders in the pilot corpus, 100 emails represented an average of 20% of all the authors' emails (8% for Germany to 34% for Zipper). It was based on this that 20% was set as the largest sample size to be extracted and attributed in the main study. Unigrams had been found to be successful in identifying the correct author of email samples, and for the main study this has been extended to six-grams, word strings of up to six words in length. The piloting process also held some very practical benefits. Firstly, it identified various issues with the Enron Email Corpus, which refined the extraction procedure and informed

the clean-up process (see Section 3.2). Also, the expansion of the experiment to use different sample sizes of authors' emails (from 20% to 2%) and to use different n-gram lengths, both led to the development of *Jangle* through various iterations and finally culminating in the version used in this study. This pilot has been considerably scaled up and has developed to form the basis of the authorship experiments reported in Chapter 6.

4.6 Chapter conclusion

This section has given an overview of the tools and the techniques used in this study. One of the main aims of this thesis is to develop a methodology (or methodologies) for authorship research that draws upon corpus, stylistic, computational and statistical approaches, that offer systematic procedures, reliable quantitative results and clear linguistic explanations. This way, we may move forward as a field, combining complementary quantitative and qualitative methodologies rather than allowing them to develop as diverging or competing approaches to the same problems. The following three chapters apply the tools and techniques described here in the analysis, attribution and profiling of authorship.

5 Identifying idiolect in the Enron Email Corpus

The nature of the theory of 'idiolect' is changing in forensic linguistics (Section 2.1). The idealised notion of idiolect as the totality of everything a speaker or writer *could* say or write at any given time (Bloch 1948: 7) is too abstract to be of any practical use to the linguist in an authorship analysis context. Instead, focus in forensic linguistics is now on measuring how distinctive an individual's language use is when tested against a relevant population of writers (Turell 2010; Turell and Gavaldà 2013; Grant 2013). Relatedly, existing research in corpus linguistics and psycholinguistics (e.g. Sinclair 1991; Nattinger and DeCarrico 1992; Wray 2002; Hoey 2005) has argued that collocational patterns and lexical co-selections are personal and idiolectal, as they are developed, stored and produced as a result of individuals' unique linguistic experiences.

Focusing on the twelve authors in the EEC12 sample (see Section 3.4.3) and using the full Enron Email Corpus as a reference set, the aim of this chapter is to use the constructs of Base Rate Knowledge (Turell 2010; Turell and Gavaldà 2013) and population-level distinctiveness (Grant 2010; 2013) to empirically test the theory that collocations and word n-grams are idiolectal of individual authors. In the development of methods for authorship attribution in forensic research, experiments are designed so that the technique is tested in extreme conditions (Chaski 2001: 4; Nini and Grant 2013: 183). With this in mind, idiolectal variation at the lexical level is sought after here within linguistic features and practices that are extremely common in the Enron population. First, distinctive ways in which the each of the EEC12 authors use the first person pronoun *I* is examined. *I* is the third most frequent word in the corpus, and an important lexical element of Enron email communication. Therefore, it presents the opportunity to analyse author-distinctive variation in the use of an extremely common word. Second, the aim is to identify the distinctive collocation profiles that different authors can have for a content word that is a central element of the shared register of the Enron employees: *deal*. Content words have consistently received criticism in authorship analysis research on the basis that they are indicative of the meaning or topic of a text, rather than author

style (e.g. Grieve 2007: 265), and this analysis counters this criticism by demonstrating the value of putting content words at the centre of discussions of idiolect. Third, idiolectal variation is examined within one of the principal speech acts for the corpus, *please*-mitigated directives. The aim with *please* is to identify the ways in which different authors express the same speech act in distinctive ways.

If the analyses here identify collocation patterns that are distinctive of individual Enron authors when tested against the population of employees from which they are drawn, this would have significant implications for forensic authorship analysis. First, they would provide empirical evidence to substantiate the theory of idiolect on which authorship attribution is founded. Second they would also show collocations and word n-grams as linguistic features, underpinned by a theory of idiolect, which could form the basis of a statistically reliable and linguistically-explainable methodology for attributing authorship.

5.1 Keywords in the Enron Email Corpus

The rationale behind choosing *I*, *deal* and *please* is that they are very frequent in the Enron Email Corpus. Stylometric approaches to authorship attribution very often focus on the most common words in the datasets in question, which predominantly comprise function words (Mosteller and Wallace 1964; Holmes et al. 2001; Burrows 2002; 2003; 2005; Argamon and Levitan 2005; Grieve 2007; Narayanan et al. 2012). However, because function words will account for the most frequent words in almost any dataset (Hoover 2009: 35), corpus linguists extract keywords from their corpus, which are 'words whose frequency is unusually high in comparison with some norm' and these provide a useful way to characterise a text or a corpus, giving a clear indication as to what the corpus is 'about' (Stubbs 2001: 129; Scott 2008b). Function words make up the top twenty most frequent words in both the Enron Email Corpus and the 464 million word *Corpus of Contemporary American English* (*COCA*). However, the results of a keyword analysis (see Section 4.1.1 for procedure) in which the Enron Email Corpus is compared against *COCA*, provide a much more useful insight into the lexis that characterises the whole corpus and the EEC12 sample (Table 7).

Table 7. Most frequent words in Enron Email Corpus and *COCA*, and Keywords in Enron Email Corpus and EEC12

Enron most frequent words			**COCA most frequent words**			**Enron Email Corpus Keywords**			**EEC12 Keywords**		
Word	**Freq.**	**%**	**Word**	**Freq.**	**%**	**Keyword**	**Freq.**	**COCA freq.**	**Keyword**	**Freq.**	**COCA freq.**
the	98,666	4.07	*the*	23,014,366	5.65	*thank(s)*	16,186	58,271	***please***	**2,176**	**43,533**
to	74,436	3.07	*and*	11,260,177	2.77	***please***	**11,061**	**43,533**	*thank(s)*	1,798	58,271
I	50,140	2.04	*of*	10,968,008	2.69	*fyi*	2,891	385	*fyi*	638	385
and	42,837	1.77	*to*	10,691,399	2.63	*attached*	3,113	12,236	*attached*	795	12,236
a	36,964	1.52	*a*	9,392,485	2.31	*fax*	2,417	6,086	***deal***	**1,183**	**87,551**
you	36,886	1.52	*in*	7,661,696	1.88	*counterparty*	1,372	8	*gas*	920	41,483
of	33,267	1.37	*that*	5,416,929	1.33	*me*	15,365	830,577	***I***	**8,590**	**4,811,110**
for	29,602	1.22	*I*	4,811,110	1.18	***I***	**50,140**	**4,811,110**	*me*	2,464	830,577
is	28,624	1.18	*it*	4,062,843	1	*am*	5,376	119,110	*thanx*	162	0
in	28,012	1.16	*is*	3,878,929	0.95	*gas*	3,577	41,483	*contract*	490	24,410
that	24,671	1.02	*for*	3,581,136	0.88	*agreement*	3,339	34,733	*review*	511	31,808
this	22,900	0.94	*you*	3,555,958	0.87	*will*	16,176	910,855	*deals*	364	11,682
on	22,061	0.91	*was*	3,074,262	0.76	*email*	1,752	4,349	*this*	3,788	1,987,129
we	22,027	0.91	*he*	3,032,348	0.74	*trading*	2,311	14,179	*am*	791	119,110
have	20,510	0.85	*on*	2,855,364	0.7	***deal***	**4,134**	**87,551**	*agreement*	492	34,733
with	18,854	0.78	*with*	2,811,606	0.69	*need*	6,677	249,664	*shall*	382	16,089
be	18,559	0.77	*as*	2,516,321	0.62	*you*	36,886	3,555,958	*it's*	211	1,431
it	17,582	0.73	*at*	2,065,603	0.51	*send*	2,827	35,340	*info*	212	1,693
me	15,365	0.63	*this*	1,987,129	0.49	*call*	5,089	149,761	*send*	464	35,340
will	15,354	0.63	*they*	1,945,940	0.48	*let*	6,266	231,727	*let*	996	231,727

Proper nouns such as company, place and people names have been removed from the Enron and EEC12 keyword lists. I *frequencies include contracted forms*.

Most of the twenty top keywords in the full Enron Email Corpus can be categorised into four types, giving a clear indication as to what the emails in the corpus are about, as well as their functions:

(i) Function words: *me, I, am, will, need* and *you*
(ii) Register-related content words: *gas, trading, deal, agreement, counterparty*
(iii) Mode of communication: *email, attached, send, call, fax*
(iv) Politeness markers: *thanks, please*

The keyword lists for the full corpus and the EEC12 sample are very similar, with twelve of the twenty most key words in the full corpus being found in the twelve-author sample (*thanks, please, FYI, attached, deal, gas, I, me, am, agreement, send, let*). This overlap lends support to the argument that EEC12 is a useful and representative sample of the full Enron corpus. In stylometric studies, relative frequency is as far as the analysis goes. Regardless of input feature sets (words, phrases, character strings, word lengths, etc.) and the algorithm or computational technique chosen, the measurement of similarity or difference between texts and authors invariably relies on the different relative frequencies with which the texts or authors in question use these features. In contrast, for corpus linguists, keywords provide a point of entry for further qualitative analysis. As such, these keyword results serve to identify three 'node' words, that is, the 'words whose co-occurrence patterns are being studied' (McEnery and Hardie 2012: 247), in the search for idiolectal collocation patterns. The three lexical items chosen are taken from three of the four categories of keywords identified. First, the first person pronoun *I* is selected given that, although it is a function word common in all English corpora (including the *COCA* reference set used here), it is still a keyword in both the full Enron corpus and the EEC12 sample, indicating that it is an especially important function word in this population. Second, *deal* is selected as it is a frequently occurring register-specific lexical item, insofar as it is sociolinguistically constrained to the context of use in the Enron community of practice, and perhaps particularly closely related to the occupational and institutional roles of individual employees (Drew and Heritage 1992: 29–32; Biber and Finegan 1994: 4; Holmes 2001: 246; Irvine 2001: 27). Third, *please* is selected as its place as the second most key word in the full corpus, and the most key word in the EEC12, suggests that *please*-mitigated requests or directives are important speech acts in Enron emails (Johnson and Wright 2014). The aim here is to identify collocational patterns and word n-

grams that are distinctive of individual authors at population level when tested against the Base Rate Knowledge of a population that uses all three of these node words extremely frequently. The assumption is that such idiolectal variation will be more difficult to identify within such commonly shared lexical items and linguistic practices than might otherwise be the case.

5.2 Idiolectal use of *I* and its collocates

Despite being a function word common in English generally, *I* is particularly common in the Enron Email Corpus. The analysis of its collocates finds that there are a number of genre- or population-level patterns which emerge in terms of its main uses by employees. In turn, some *I*-initial lexico-grammatical patterns and word n-grams appear to be distinctive of individual authors in the corpus, providing evidence for idiolectal uses of *I* within this population. This underlines the value of moving beyond the relative frequency results relied upon in stylometric analyses and towards a collocational approach to analysing style.

I is the third most common word in the Enron Email Corpus, accounting for 50,140 of all 2,462,151 (2.04%) tokens, compared with only 1.18% of all words in *COCA*. This 2.04% represents the Base Rate Knowledge of how frequent *I* is in this population of writers. This relatively high frequency of *I* in this corpus is likely to be due to the interactional and personal nature of workplace emails. Titak and Roberson (2013: 247–8), also analysing the Enron corpus, found that in comparison with other online genres, workplace emails 'involve a more personal and narrative style' and 'rely on specific features such as personal pronouns and verbs – in common with most spoken, highly interactive communication'. It appears, however, that Enron email discourse differs in its distribution of *I* from other workplaces and types of institutional communication, particularly spoken discourse. Poncini (2004: 119), for example, studied spoken multicultural business meetings, and argued that participants achieve and facilitate task-oriented discourse through the use of personal pronouns. However, she found that *I* was far less common than *we* and *you* in her data. Similarly, Planken (2005: 396) studied the use of personal pronouns by professional business negotiators and students of international business management. She found *we* to be an important device, used either as an indicator of solidarity and co-cooperativeness or of professional distance. In contrast, direct references to *I* in speech were described as being self-orientated, 'highly subjective'

and potentially face-threatening in negotiations. Vine (2004: 96) investigates spoken discourse within four government workplaces in New Zealand and finds *we* to be most important in the mitigation and modification of implicit directives (or 'control acts'). Therefore, it might be that *I* is not only more salient in the Enron Email Corpus when compared with American English generally, but it is also more common than in other workplaces and types of institutional discourse.

Immediately, differences can be identified between the authors in terms of how frequently they use *I*. Using the log-likelihood statistic (described in Section 4.4.1), it is found that although ten of the twelve authors use *I* more frequently than the 2.04% Base Rate for the corpus, the difference in frequency from this Base Rate is only statistically significant (at the $p<0.0001$ level) for three authors—Germany, Lavorato and Dorland (Table 8).

Table 8. Frequency of *I* across the twelve EEC12 authors

Author	**Tokens**	***I* freq.**	**% of tokens**
Dorland	14,565	572	3.93%*
Germany	77,597	2,703	3.48%*
Lavorato	24,677	744	3.01%*
Derrick	5,902	156	2.64%
Kaminski	52,992	1,393	2.63%
Zipper	7,393	190	2.57%
Haedicke	18,540	465	2.51%
Arnold	26,283	628	2.39%
Farmer	24,502	575	2.35%
Nemec	50,211	1,117	2.22%
Allen	14,579	274	1.88%
Steffes	35,284	617	1.75%

*statistically significant difference from Base Rate at p <0.0001 using log-likelihood statistic

Their authorial styles, therefore, are distinctive in that they use *I* more frequently than is expected in the population from which they have been taken. It is at this comparison of relative frequencies of function words such as *I* where stylometric approaches end. Such results can only take the analyst so far. Although Dorland, Germany and Lavorato have been identified as high frequency users of *I*, nothing has been revealed in terms of the ways in which they use *I* that can serve to distinguish their idiolects from each other or the rest of the Enron population. It is

here where analysis can extend to collocational patterns with *I*, and measure the extent to which these can be considered idiolectal.

Table 9 lists R1 collocates that account for more than 1% of *I* instances in the Enron corpus and the EEC12 sample.

Table 9. Most frequent *I* collocates in the Enron Corpus and the EEC12 sample

	Enron Corpus				EEC12			
	R1	**Authors**	**Freq.**	**% of *I***	**R1**	**Authors**	**Freq.**	**% of *I***
1	*have*	156	4,528	9.03	*am*	12	693	8.26
2	*am*	146	4,387	8.75	*have*	12	653	7.78
3	***will***	**152**	**4,370**	**8.72**	***think***	**12**	**611**	**7.28**
4	***think***	**131**	**2,815**	**5.61**	***will***	**12**	**597**	**7.11**
5	***would***	**142**	**2,607**	**5.2**	***would***	**12**	**519**	**6.18**
6	*don't*	115	1,955	3.9	*don't*	11	379	4.52
7	***can***	**136**	**1,680**	**3.35**	*shall*	1	317	3.78
8	*was*	140	1,334	2.66	***need***	**12**	**220**	**2.62**
9	***need***	**129**	**1,284**	**2.56**	*just*	11	209	2.49
10	*just*	120	1,038	2.07	***can***	**12**	**199**	**2.37**
11	***know***	**142**	**920**	**1.83**	*was*	12	174	2.07
12	*had*	113	702	1.40	***want***	**12**	**136**	**1.62**
13	***hope***	**101**	**668**	**1.33**	***know***	**12**	**110**	**1.31**
14	*do*	137	643	1.28	***believe***	**11**	**94**	**1.12**
15	***believe***	**105**	**640**	**1.28**	*had*	12	87	1.04
16	***thought***	**99**	**633**	**1.26**	*didn't*	9	84	1.00
17	***want***	**113**	**509**	**1.02**				
18	*didn't*	76	505	1.01				
		Total	**31,218**	**62.26**		**Total**	**5,082**	**60.55**

Again, as with the keywords, fifteen of the top eighteen most frequent collocates in the corpus are also in the most frequent collocates in EEC12. Taken together, these top collocates account for as many as 62.26% of all *I* instances in the Enron Email Corpus; they offer a Base Rate Knowledge of what *I* most frequently collocates with in the Enron population. Two major patterns that emerge in this regard are that *I* very frequently collocates with mental processes *(think, need, know, hope, believe, thought, want)* and modal auxiliary verbs (*will, would, can*). Furthermore, there is an interesting overlap between these two types of collocates. Of the 2,607 instances of *I would* in the Enron Email Corpus, *I would like* occurs 1,023 times. Aggregated, mental process and modal verbs account for 32.16% of *I* instances: over half of the

most frequent collocates and a third of total collocates in the corpus. A closer qualitative investigation of these patterns can reveal author-distinctive usage.

5.2.1 *I will*

In a workplace context, modal verbs 'represent a major resource for speakers in expressing interpersonal meanings, whether they are used for transactional or relational purposes' (Koester 2006: 77). In terms of modal verb collocates of *I*, the twelve EEC12 authors can be distinguished from each other in three ways: (i) the amount they modalise; (ii) the specific modal verbs they use; (iii) the different ways they use the same modal verb. As the most frequent *I* + modal verb collocation, *I will* is a good example to use to investigate idiolect in this way.

In relation to (i), the amount with which authors modalise, Table 10 shows all eleven modal verbs that *I* collocates with in the Enron Email Corpus, and compares the frequencies with which the EEC12 authors use them. These eleven modal verbs collocate with *I* with a cumulative frequency of 10,331, accounting for 20.6% of all 50,140 occurrences of *I* in the Enron Corpus. This gives a Base Rate Knowledge of how frequently *I* is followed by a modal verb in the population. Compared against this Base Rate Knowledge, eight of the twelve authors—Kaminski, Derrick, Haedicke, Allen, Farmer, Nemec, Zipper, Lavorato—can be said to collocate *I* with a modal verb more frequently than can be considered the norm, as they all do so in more than 20.6% of *I* instances. However, based on log-likelihood tests, only Kaminski uses *I* + modal verb significantly more frequently than the Base Rate for the corpus. As such, the frequency with which he co-selects *I* and a modal verb marks his writing style as being distinctive against the population from which he is taken. By the same token, the opposite is the case for Germany and Arnold, who collocate *I* with modal verbs significantly *less* frequently than the Base Rate for the corpus.

In relation to (ii), authors' preferences for different modal verbs with *I*, another distinctive element of Kaminski's data is that he is the only one of the twelve EEC12 authors who uses *I shall* (Figure 15).

Table 10. *I* + MODAL collocations across the EEC12 authors

Author	***I***	***will***	***would***	***shall***	***can***	***may***	***could***	***should***	***have to***	***might***	***must***	**Total**
	Base Rate	(8.72%)	(5.20%)	(0.63%)	(3.35%)	(0.57%)	(0.74%)	(0.68%)	(0.40%)	(0.21%)	(0.09%)	(20.60%)
Kaminski	1392	23 (1.65)*	84 (6.03)	**317 (22.77)***	60 (4.31)	10 (0.72)	26 (1.87)*	6 (0.43)	7 (0.50)	1 (0.07)		**534 (38.36)***
Derrick	147	**31 (21.09)***	11 (7.48)			1 (0.68)	1 (0.68)					44 (29.93)
Haedicke	465	55 (11.83)	54 (11.61)*		13 (2.80)	2 (0.43)	1 (0.22)	3 (0.65)	2 (0.43)			130 (27.96)
Allen	272	23 (8.46)	30 (11.03)		9 (3.31)	2 (0.74)	2 (0.74)			1 (0.37)		67 (24.63)
Farmer	530	60 (11.32)	27 (5.09)		21 (3.96)	4 (0.75)	4 (0.75)	2 (0.38)	1 (0.19)			119 (22.45)
Nemec	1104	**159 (14.4)***	51 (4.62)		24 (2.17)	1 (0.09)	2 (0.18)	4 (0.36)	1 (0.09)	1 (0.09)	1 (0.09)	244 (22.1)
Zipper	177	18 (10.17)	16 (9.04)		2 (1.13)		2 (1.13)	1 (0.56)				39 (22.03)
Lavorato	598	30 (5.02)	69 (11.54)*		13 (2.17)	3 (0.50)	6 (1.00)	2 (0.33)	2 (0.33)	1 (0.17)		126 (21.07)
Dorland	520	42 (8.08)	29 (5.58)		14 (2.69)	1 (0.19)		5 (0.96)	4 (0.77)	1 (0.19)		96 (18.46)
Steffes	491	20 (4.07)	37 (7.54)		10 (2.04)	6 (1.22)	1 (0.20)	3 (0.61)	1 (0.20)			78 (15.89)
Germany	2218	122 (5.5)*	88 (3.97)		27 (1.22)*	34 (1.53)*	10 (0.45)	17 (0.77)	3 (0.14)	5 (0.23)	4 (0.18)	310 (13.98)*
Arnold	480	14 (2.92)*	23 (4.79)		7 (1.46)	4 (0.83)	3 (0.63)	2 (0.42)	3 (0.63)	1 (0.21)		57 (11.88)*
Total	**8394**	**597**	**519**	**317**	**200**	**68**	**58**	**45**	**24**	**11**	**5**	

*statistically significant difference from Base Rate at $p < 0.0001$ using log-likelihood statistic

Figure 15. 10 of 317 concordance lines for *I shall* in Kaminski's data

N	Concordance
1	review the draft from the point of view of our PR policy. I shall read it as well. Vince <Message-ID: > Rick, I shall ask
2	policy. I shall read it as well. Vince <Message-ID: > Rick, I shall ask my assistant to schedule a meeting early next
3	Weekend at Stanford. Please, send me the slides anyway. I shall help Tanya to prepare her presentation. Vince
4	week. Vince <Message-ID: > Elsa, Thanks for the invitation. I shall be glad to join you. Please, send me the details. Vince
5	between upsetting Neil or Jeffs (Shankman and Skilling), I shall choose Neil. Vince Enron North America Corp.
6	on trade ideas, over which we have no control long-term. I shall talk to Rick Buy and David Port about setting up more
7	the building. Vince <Message-ID: > Shalesh, A good idea. I shall forward it to the AG traders. Vince <Message-ID: >
8	L1 visa anyway and we decided to go ahead an arrange it. I shall also write to him and explain the confusion. Also, if I
9	Vince <Message-ID: > University of Texas at Austin Joe, I shall probably ask Tanya to attend. It coincides with
10	disciplines (math, physics) or computer programming. I shall send your resume to some other units in the company

Kaminski uses most often uses *I shall* to make epistemic commitments, such as *I shall call* which he uses twenty times, *I shall ask* which he uses eighteen times, *I shall send* which occurs seventeen times, and *I shall talk* fifteen times (Examples 1–4).

Example 1

<From: kaminski@enron.com>
<To: paul.quilkey@enron.com>
<Subject: RE: Weather>

Paul,
No problem. **I shall call** you this evening.
Vince

Example 2

<From: vince.kaminski@enron.com>
<To: anjam.ahmad@enron.com>
<Subject: Re: Video Conference for Interview: Stig Faltinsen>
<Cc: vince.kaminski@enron.com, shirley.crenshaw@enron.com>
<Bcc: vince.kaminski@enron.com, shirley.crenshaw@enron.com>

Anjam,
Sorry, I am busy on Thursday.
I shall ask Shirley to contact you. Friday 9:30 to 10:30 my time would work.
Vince

Example 3

<From: j.kaminski@enron.com>
<To: remi.collonges@enron.com>
<Subject: FW: A resume>

Remi,
FYI. **I shall send** you the summary of action items later today.
Vince

Example 4

<From: vince.kaminski@enron.com>
<To: richard.causey@enron.com>
<Subject: UofT>
<Cc: vince.kaminski@enron.com>
<Bcc: vince.kaminski@enron.com>

Rick,
Thanks for your message. **I shall talk** to Greg Whalley about his participation.
Vince

In addition to these, he also uses *I shall* to mark futurity, especially in *I shall be,* which he uses in reference to where he will be or what he will be doing on a certain date or time, as in 44 instances (Examples 5–7) or in the phrase *I shall be glad* which he does 33 times (Examples 8–9). Kaminski is the only author of all 176 in the Enron Email Corpus to use the bigram *I shall*, which he uses 317 times.

Example 5

<From: vince.kaminski@enron.com>
<To: eugenio.perez@enron.com>
<Subject: Re: Dinner>

Eugenio,
Thanks for the invitation. Can we reschedule the dinner?
I shall be in San Antonio on Friday.
Vince

Example 6

<From: vince.kaminski@enron.com>
<To: dale.nesbitt@marketpointinc.com>
<Subject: Re: Follow up>

Dale,
I have passed on the information you gave me but you have to realize that I acted just as a go-between. I shall be glad to remind our new business unit about your proposal. I have already asked a few times. **I shall be traveling** this week, Wed - Fri. I shall be back in the office on Monday.
Vince

Example 7

<From: j.kaminski@enron.com>
<To: eydie.corneiro@enron.com>
<Subject: Declined: Mtg w/Vince Kaminski>

Eydie,
I shall be out on Monday, July 9. What about the following Tuesday?
Vince

Example 8

<From: vince.kaminski@enron.com>
<To: louise.kitchen@enron.com>
<Subject: Spring 2001 Energy Finance Conference Participation>

Louise,
Would you consider being a keynote speaker at this conference (Feb 22 evening)? The conference will be held in Austin. We have a very good relationship with UT and we are helping them to organize, this conference. **I shall be glad** to provide you more information about the event.
Vince

Example 9

<From: kaminski@enron.com>
<To: traci.warner@enron.com>
<Subject: CMU>

Traci,
I shall be glad to work with your team at Carnegie Mellon.

In *I shall be glad*, Kaminski appears to be using an alternative to *I would be glad/happy to*, which he uses only three times in his data. In all other instances, he is showing a preference for *I shall* over *I will*, which would be a suitable means of expressing epistemic commitment and futurity. This is reflected in the finding in Table 10 above which shows that in comparison with the 22.77% of *I* instances he collocates with *shall*, Kaminski co-selects *will* with only 1.65%. He is the least frequent user of *I will* of the EEC12 authors, and uses it statistically significantly *less* frequently than the 8.72% Base Rate for the corpus. On the basis that *I shall* is unique to Kaminski when tested against a 'relevant population' in the 'the same speech community' (Turell and Gavaldà 2013: 499) from which he is taken, this bigram can be said to be a distinctive manifestation of his idiolect. Beyond the Enron Email Corpus, *I shall* is far rarer in American English generally than *I will*. In *COCA*, *I shall* is found 3,468 times, compared with *I will* which occurs 36,897 times, eleven times as frequently. Similarly, a Google search for *'I shall*' (in quotation marks to keep search terms together) returns 15.4 million results, which is only 5% of the 294 million search results returned for '*I will'*. Crystal (1986: 42) discusses the debate surrounding *will* and *shall* which began in the early part of the seventeenth century. Today, however, contemporary reports such as Williams (2013) and Tagliamonte et al. (2014: 78–9) tell us that in modern English worldwide, *shall* has receded to restricted, formulaic and legal usage. Therefore, in

addition to being unique to Kaminski at population level within the Enron Email Corpus, the overwhelming preference for *I shall* over *I will* is distinctive beyond this relevant population and in English generally. Kaminski's retention of *I shall* in a language in which it is otherwise declining may be attributable to the fact that he is Polish (*New York Times* 2006) and a likely to be a non-native speaker of English. A video interview with him now as Professor in the Practice of Energy Management validates this assumption (YouTube 2012). It might be that in learning the language, Kaminski was taught that *shall* is to be used for the simple future in declarative sentences with a first person subject, a rule which is still taught today (Hoye 2014: 120; *The Economist* 2011; *BBC World Service* 2011. See also http://www.oxforddictionaries.com/words/shall-or-will). This is speculation, of course, but the reality is that some aspect of Kaminski's unique language learning, contact and experience has led to the *I shall* collocation being preferred to *I will* in his idiolect, which it is not the case for any of the other 175 Enron employees. It might also be that his frequent uses of *shall* account for Kaminski having a higher modalisation rate than the other EEC12 authors and the Base Rate for the full Enron corpus.

With the exception of Kaminski's distinctive *shall*, *will* is the modal verb which collocates the most with *I* (Table 10). Of the EEC12 authors, two are distinguished as particularly frequent users of this collocation. Derrick (21.09%) and Nemec (14.40%), both of whom are lawyers, use it significantly more frequently (as a proportion of *I* instances in their data) than the Base Rate for the corpus. With regard to (iii), the different ways authors use the same modal verb, Derrick and Nemec exhibit longer *I will* sequences which distinguish them from each other and from the Enron population as a whole. Unless stated otherwise, in this chapter a lexical string is considered as being 'distinctive' of an author if they are the only person in the Enron Email Corpus to use it more than once (see Wright [2013: 53] which also used this approach). The verbs which Derrick uses *I will* to modalise on more than one occasion are *not be* (9), *be* (6), *attend* (6) *ask* (3), and *support* (3). Many of these are part of longer word n-grams that Derrick recurrently uses, such as *I will not be in*, *I will not be able to attend*, and *I will support your* (Figure 16). Derrick uses *I will not be in* on two occasions (lines 16 and 17), both of which are in almost identical emails he sends to his recipients to inform them that he will not be

able to accept their invitation to an event. On both occasions, he is referring to a city:

Figure 16. 27 of 33 *I will* concordance lines in Derrick's data

N	Concordance
1	> Dave, thank you for the suggestion. I will ask V&E to draft a memo. All the
2	<Message-ID: > Dave, I am out of town. I will ask Stephanie to contact your
3	All the best. Jim <Message-ID: > Rex, I will ask him to include you on the list.
4	> I agree. <Message-ID: > Steph,I will attend this. <Message-ID: >
5	and Tuesday morning. <Message-ID: > I will attend the Monday evening meeting
6	you. Jim <Message-ID: > Vanessa, I will attend. Jim <Message-ID: >
7	that Ken received. Jim <Message-ID: > I will attend. Jim Jim Derrick
8	to that message. Jim <Message-ID: > I will attend in person. Jim <Message-ID:
9	> I have replied to this, indicating I will attend both Monday night and
10	and communicating via Blackberry. I will be the office tomorrow. Give me a
11	signature. Thank you. <Message-ID: > I will be pleased to meet with Kersten
12	Anthony Cann on Monday, I regret that I will be on the east coast on that day. I
13	get it paid. Jim <Message-ID: > David, I will be faxing this to you in a few
14	in connection with those meetings, I will be departing for San Antonio.
15	> Please call Beal's and tell them I Will be at Inwood Manor Around 1:15.
16	you for the invitation. Unfortunately, I will not be in Houston on Friday night
17	to the June 14 reception. Unfortunately, I will not be in London on that day and
18	Thank you. <Message-ID: > Steph, I will not be able to attend. I will need to
19	Southwestern Legal Foundation that I will not be able to attend the November
20	Thank you. <Message-ID: > Steph, I will not be able to attend this meeting.
21	because of the press of business, I will not be able to attend the Silverado
22	you. <Message-ID: > Meeting Steph, I will not be able to attend this meeting.
23	you. <Message-ID: > Please reply that I will not be able to attend. Jim Derrick
24	conflicts with the Legal PRC meeting, I will not be able to attend. Thank you for
25	ID: > FYR <Message-ID: > Gail, I will support your recommendation. Jim
26	satisfied that he can do the job well, I will support your recommendation; and
27	age-ID: > FYI <Message-ID: > I will support your and Chris's decision.

Example 10

<From: james.derrick@enron.com>
<To: cindy.olson@enron.com>
<Subject: RE: Invitation>

Thank you for the invitation. Unfortunately, **I will not be in Houston** on Friday night and thus will not be able to attend. Jim Derrick

Example 11

<From: james.derrick@enron.com>
<To: maria.scudder@linklaters.com>
<Subject: The Flight Gallery>

Thank you for the invitation to the June 14 reception.
Unfortunately, **I will not be in London** on that day and thus will be
unable to attend. Jim Derrick of Enron in Houston, Texas USA

This seemingly unremarkable collocational string is used more than once by only one of the other 175 authors in the Enron corpus, and so can be considered very distinctive of Derrick. Sara Shackleton uses *I will not be in* four times, but each time refers to *the office*, rather than cities. There is a similar case with *I will not be able to attend*. Derrick uses this string seven times (lines 18–24), three of which he ends with a full stop, and four to which he adds *this meeting*, *the Silverado meeting* or *the November meeting*. In the rest of the Enron Corpus, *I will not be able to attend* is used more than once by seven authors, none of whom extend the clause with *this* or *the* as Derrick does. Finally, on three occasions Derrick writes *I will support your* (lines 25–27), which is not found at all in the rest of the Enron corpus. A Google search for the longer string *I will support your recommendation* returns only four results from the billions of documents searched, underlining its rarity and distinctiveness, and none of these are a single punctuated sentence as they are for Derrick. All of these word n-grams, although low in frequency, are distinctive of Derrick at population level in the Enron Corpus and beyond. On this basis it can be argued that they are identifiable segments of his idiolect. Sapir (1927: 903–4) argues that: 'there is always an individual method, however poorly developed, of arranging words into groups and of working these up into larger units'. In turn, Wray (2002: 174) argues that individual differences in the use of formulaic sequences are linked to variation in the interactional experiences of language users. Similarly, Hoey (2005: 9) claims that every time a word is used or encountered by an individual, this experience strengthens the association between this word and its immediate collocational context. In the same way as *I shall* for Kaminski, then, it appears as though these collocational strings are manifestations of Derrick's unique language experiences. In turn, their reproduction in his workplace emails distinguishes his authorial style from that of the other employees within the same linguistic community.

These distinctive patterns contrast with Nemec's, the second most frequent user of *I will* in the EEC12 sample. The verbs which he most frequently co-selects with *I will* are *be* (34), *forward* (18), *prepare* (9), *let* (8), and *review* (7). Like Derrick, Nemec is a lawyer in Enron (specifically the attorney for Enron Capital and Trade Resources) a role which for him, and for other lawyers in the company, involved collaborative drafting, revising and redrafting of legal documents and agreements. Devitt (1991: 336–7) studied the range of written genres used by tax accountants and argues that texts are both the 'resources' and the 'products' of their profession. They are resources in that they provide the authority for what the accountants write, and they are the products as they charge fees for the documents they construct. This is a similar situation to the lawyers in Enron; state and federal legislation are the texts that guide and govern their work, legal agreements are the product of their work, and emails are the texts they use to produce them. In turn, a number of Nemec's recurring *I will* sequences are related to this occupational practice (Figure 17).

Figure 17. 20 of 159 concordance lines for *I will* in Nemec's data

N	Concordance
1	with revisions. Please review. I will be forwarding the coal handling
2	Attached is a redline with my revisions. I will be forwarding two hard copies of the
3	Please take one last look. If OK I will forward the ECS purchase to Porter
4	and the anticipated start date is May 1. I will forward the filing TW made wrt to
5	agreement. Please provide comments. I will forward the revised Huber
6	attached is the revised CSA. I will forward the revised Lease and O&M
7	call me if you have any questions and I will forward the Rio Nogales later today.
8	Upon receipt of comments, I will forward to Bill Howell.
9	county yet. As soons as I receive it, I will forward to John. <Message-ID: >
10	drive and WT1. Please review and if OK I will forward to Susan for her review.
11	which needs to be provided. If OK I will forward to TW. <Message-ID: >
12	<Message-ID: > Thanks for the info. I will prepare the amendment.
13	you two credit tickets for these entities. I will prepare the confirms for the Nevada
14	you. No problem. <Message-ID: > I will prepare the documents. Attached is
15	forward to Oneok for their review. I will prepare the executable form of
16	> The final sign off from Reliant legal. I will prepare the executables and get
17	terms. Please provide this info, and I will prepare the tasking letter.
18	which I modified to fit this context. I will prepare the tasking letter based on
19	for the delay. <Message-ID: > I'm back. I will review and comment on their
20	a due diligence the last couple of days. I will review and get back to you
21	the wrong file. <Message-ID: > I will review and get back to you by no

Some of the other sequences Nemec uses such as *I will be out, I will be in,* and *I will call* are extremely common in the Enron Corpus, being used by 67, 56 and 42 other authors respectively. These more register-related verbs and n-grams, however, are far more distinctive of Nemec. For example, *I will prepare the* occurs seven times in Nemec's data (lines 12–18), as he refers to his preparing of *executables, tasking letters, confirms* and *documents*. This is a clear example of an n-gram relating to the drafting process of legal documents (Examples 12–13).

Example 12

<From: gerald.nemec@enron.com>
<To: barry.tycholiz@enron.com>
<Subject: Re: FW: Producer Payment Summary>

I will prepare the documents. Attached is the contract form that Crestone forwarded for their payment. This would be a schedule to an existing Master Services Agreement in place between ENA and Crestone.

Example 13

<From: gerald.nemec@enron.com>
<To: nick.cocavessis@enron.com>
<Subject: PCS Pro. Services Agreement>

Nick, Attached is the form of services agreement. The form is a little different than Steve Van Hooser's original form. This a form that we used in Wyoming which I modified to fit this context. **I will prepare the** tasking letter based on your scope of work and forward shortly. Let me know if you have any questions.

This *I will prepare the* sequence is only found three times in the rest of the Enron Corpus, two of which are used by another lawyer, Debra Perlingiere. In a population of 176 Enron employees, many of whom are lawyers or involved with the legal dealings of the corporation, only two people use this n-gram. As such, it is highly distinctive of Nemec's idiolect. Similarly, Nemec has a number of n-grams which refer to forwarding the drafted and revised documents onwards. He uses *I will be forwarding* twice (lines 1–2), *I will forward the* five times (lines 3–7) and *I will forward to* five times (lines 8–11). All three of these are very rare in the rest of the Enron corpus. First, *I will forward the* is used a total of thirteen times by nine different authors, only three of whom use it more than once: Tana Jones, Kay Mann and Debra Perlingiere – all lawyers. Second, *I will forward to*, in which the direct object is ellipted is rarer, being found eleven times by nine authors, only two of whom use it more than once: Elizabeth Sager and Dan Hyvl. Third, *I will be*

forwarding is only found twice in the rest of the corpus, used by two different authors, neither of whom repeat it. Finally, Nemec uses *I will review and* three times (lines 19–21) in such a way that he commits to reviewing a document and subsequently contacting the recipient (Example 14–15). This four-gram is not used by any of the other 175 authors in the Enron corpus.

Example 14

<From: gerald.nemec@enron.com>
<To: michael.legler@enron.com>
<Subject: Re: Latest WIC Supplement>

Yes, I did receive it. I have been out of the office on a due diligence the last couple of days. **I will review and** get back to you tommorrow. Thanks.

Example 15

<Date: Mon, 10 Jul 2000 03:07:00 -0700 (PDT)>
<From: gerald.nemec@enron.com>
<To: stephanie.miller@enron.com>
<Subject: Re: Williston Basin PA>

I will review and get back to you by no later than tommorrow.

Nemec is writing within a community of practice which includes a number of people who are sociolinguistically similar to him in terms of training, experience and expertise. In addition, he is engaging in activities that are common to many other employees in the corpus. *Review*, for example, is found 1,969 times in the corpus and is used by 103 authors, while *forward* occurs 1,914 times across the emails of 123 authors. Nevertheless, Nemec produces job-related n-grams that are either very rare in the rest of the corpus, or not found at all. It can inferred, therefore, that aspects of his storage of linguistic units and lexical primings have been affected in a way that is in some way different to that of his colleagues. The result is his production of word n-grams that are distinctive of him when tested against the population from which he is taken.

Employing the concepts of Base Rate Knowledge and population-level distinctiveness, an analysis of *I will* has identified various types of potentially idiolectal evidence in the emails of Enron employees. Based on relative frequencies, it was found that Kaminski collocates *I* with modal verbs more frequently than is expected in the Enron community of practice. He is also the only one of the 176 authors who uses the receding *I shall* collocation, a collocation he prefers over the

far more common *I will*. Derrick and Nemec were identified as the most frequent *I will* users in EEC12, as they both use it more frequently than the Base Rate for the corpus. In turn, a closer qualitative examination of the ways in which these two authors use *I will* has revealed a number of potentially idiolectal collocations and word n-grams. Each of these types of distinctiveness, it is argued here, is a result of the unique linguistic experiences and interactions these authors have had throughout their lives.

5.2.2 *I think*

'Mental' processes are concerned with our experience of the world of our own consciousness, 'processes of feeling, wanting, thinking and seeing' (Halliday and Matthiessen 2004: 197, 207). One mental process which has received a lot of attention is *think*, particularly with a first person subject (e.g. Aijmer 1997; Karkkainen 2003; Kaltenbock 2009; Fetzer 2008). *Think* is the mental process which collocates most frequently with *I* in the Enron corpus, and is the fourth most frequent *I* collocate overall (after *have, am* and *will*).

Fetzer (2008: 389 and 2014: 68) in her analysis of *I think* in political discourse, argues that *I think* no longer expresses solely the cognitive disposition of the speaker, but rather his or her attitude and epistemic commitment toward the underlined proposition, and invites the listener or reader to adopt this perspective. Zhang (2014) outlines a number of functions for *I think.* On the one hand, its function is of 'an epistemic nature, expressing either evaluation or emphasis to assert the speaker's credibility and authority (Zhang 2014: 253). On the other hand, Zhang (2014: 226) highlights that it can be used to tone *down* assertiveness and authority, mitigating otherwise face-threatening acts. In workplace discourse, this multi-functional nature of *I think* is well-documented. Koester (2006: 81–2) for example, notes that it can be used to express a relative degree of commitment (e.g. *I don't think she's there*), as well as expressing opinions (e.g. *I think it's a good idea*). She states that the latter of these can be considered part of deontic modality, and these often 'occur in conversations involving decision-making' where speakers express their 'opinions and judgements in discussing and evaluating events', for example in *I think it looks better without it*. Handford and Matous (2011: 94–5) find that in their data of on-site interactions in the international construction industry, *I think* is the fourth most frequently occurring cluster. They describe this cluster as a 'hedging expression', without elaboration as to whether it hedges epistemic

commitment or mitigates deontic expressions. They do comment, however, that its frequency 'suggests that face needs are a concern of the speaker'. Vine (2004: 101) aligns with this argument, claiming that in her workplace data '*I think* acts as a hedge, softening the head act', in such a way that users are not 'taking full responsibility' for their utterance, as in the directive*: I think you're gonna need to do this*.

These various uses of *I think* in a workplace context vary from author to author in the Enron corpus, with different people using it in different ways with different functions. Most crucially for the identification of idiolectal variation, however, is that these uses and functions are expressed in distinctive ways by individual authors. To exemplify this, focus will be on Haedicke, Arnold and Steffes as the most frequent users of *I think* in EEC12. These three authors are distinguished from the others in EEC12 on the basis that they co-select *think* with *I* significantly more frequently than the 5.61% Base Rate norm for the corpus (Table 11). These three authors all use *I think* to hedge a commitment or certainty (Example 16–17), express an opinion (Example 18–19) and to mitigate directives (Example 20–21).

Example 16

<From: jim.steffes@enron.com>
<To: j.noske@enron.com>
<Subject: FW: Another Budget Meeting>

I'm taking Sept 3 off - **I think it's Labor Day**. Please also note that I am taking off late in September.

Example 17

<From: mark.haedicke@enron.com>
<To: julia.murray@enron.com>
<Subject: Re: Summer Vacation Plans>

I will not be back from my vacation until August 2. Is it possible to delay your vacation for two days. If not, **I think it will be ok**. Mark

Example 18

<Date: Tue, 2 Oct 2001 09:47:39 -0700 (PDT)>
<From: john.arnold@enron.com>
<To: jennifer.fraser@enron.com>
<Subject: RE: changes to curve?>

[...] In this case **I think 100 is necessary** because once a strike has open interest, we must continue to support it.

Table 11. *I* + mental process collocations across the EEC12 authors

	I think	***I need***	***I know***	***I hope***	***I believe***	***I want***	***I thought***	***I guess***	***I assume***	***I agree***
(Base Rate)	**(5.61%)**	**(2.56%)**	**(1.83%)**	**(1.33%)**	**(1.28%)**	**(1.02%)**	**(1.26%)**	**(0.69%)**	**(0.44%)**	**(0.50%)**
Haedicke	**70 (15.05)***	6 (1.29)	2 (0.43)	2 (0.43)	6 (1.29)	22 (4.73)*	1 (0.22)	1 (0.22)	1 (0.22)	9 (1.94)
Steffes	**64 (13.03)***	11 (2.24)	12 (2.44)	7 (1.43)	5 (1.02)	23 (4.68)*	6 (1.22)	2 (0.41)	15 (3.05)*	15 (3.05)*
Arnold	**56 (11.67)***	9 (1.88)	3 (0.63)	4 (0.83)	3 (0.63)	5 (1.04)	4 (0.83)	7 (1.46)	5 (1.04)	1 (0.21)
Zipper	20 (11.30)	5 (2.82)	4 (2.26)		1 (0.56)		3 (1.69)			4 (2.26)
Lavorato	57 (9.53)	20 (3.34)	12 (2.01)	4 (0.67)		16 (2.68)	22 (3.68)*	11 (1.84)	8 (1.34)	7 (1.17)
Kaminski	97 (6.97)	8 (0.57)*	6 (0.43)*	26 (1.87)	0 (0.00)	17 (1.22)		2 (0.14)	11 (0.79)	13 (0.93)
Germany	146 (6.58)	72 (3.25)	36 (1.62)	19 (0.86)	52 (2.34)	24 (1.08)	28 (1.26)	19 (0.86)	12 (0.54)	2 (0.09)
Dorland	28 (5.38)	23 (4.42)	5 (0.96)	2 (0.38)	0 (0.00)	10 (1.92)	4 (0.77)	3 (0.58)		
Allen	13 (4.78)	15 (5.51)	6 (2.21)	2 (0.74)	7 (2.57)	3 (1.1)	3 (1.10)	1 (0.37)	2 (0.74)	
Nemec	47 (4.26)	29 (2.63)	18 (1.63)	1 (0.09)*	5 (0.45)	11 (1)	6 (0.54)	2 (0.18)	4 (0.36)	6 (0.54)
Farmer	11 (2.08)*	22 (4.15)	5 (0.94)	2 (0.38)	11 (2.08)	4 (0.75)	5 (0.94)	2 (0.38)	1 (0.19)	2 (0.38)
Derrick	2 (1.36)		1 (0.68)	1 (0.68)	4 (2.72)	1 (0.68)			5 (3.40)	4 (2.72)

*statistically significant difference from Base Rate at $p < 0.0001$ using log-likelihood statistic

Example 19

<From: jim.steffes@enron.com>
<To: l.nicolay@enron.com, janel.guerrero@enron.com, alan.comnes@enron.com,>
<sarah.novosel@enron.com>
<Subject: RE: CN comments on well designed RTO>

I think that Christi's write up is great. I've only changed two things. If anyone has any more changes, please get to Janel ASAP.
Jim

Example 20

<From: mark.haedicke@enron.com>
<To: daniel.rogers@enron.com>
<Subject: Re: Middle East Legal Support>

Dan:
I think you should add a sentence that you will continue to be directly involved in LNG matters including the contracts for Dabhol and that additional resources are being developed in London both inside and outside.
Mark

Example 21

<Date: Mon, 14 May 2001 11:08:00 -0700 (PDT)>
<From: mark.haedicke@enron.com>
<To: david.oxley@enron.com>
<Subject: Re: Keohane>

I think we will need to do something for Peter. I would be interested in HRs view on how much. Mark

The differences between these types of *I think* uses are not always clear-cut. For instance, in Example 17 the highlighted *I think it will be OK* could be considered as a hedged expression of opinion, and in Example 20 the highlighted utterance could be considered both an indirect directive, mitigated by *I think,* and an expression of opinion. Investigating the specific differences between Arnold, Haedicke and Steffes, distinctions between the authors' *I think* patterns and preferences become clearer. To start with Arnold, no particularly distinctive patterns emerge. His 56 instances of *I think* are followed by 33 different words. Most of these he uses only once, and those which he uses more than once include *I think the* (7 times), *I think it's* (4), *I think I* (3) and *I think they* (3) and *I think we* (3) (Figure 18). None of these are part of longer recurring strings, and all are very common in the Enron corpus. *I think the,* for example, is used 212 times by 66 authors, and *I think I* is used 206 times by 73 authors.

Figure 18. 20 of 56 *I think* concordance lines in Arnold's data

N	Concordance
1	on the floor so we have to harass them when we do. I think I am a couple of the "we are not" attributes in your
2	<Message-ID: > relaxed? how was it? <Message-ID: > i think i had better weather in massachusetts than you had
3	Mike 4:30-5:00 Thanks, John <Message-ID: > Russell: I think I should give you a little background on small
4	I'm saying there is a much better chance than that. i think it's 2:1 <Message-ID: > 2.75 ... but yea
5	to 2.8. whther that price level is 425 or 725 is arguable. i think it's close to here. but when we get to november and
6	If you want to buy this strip back in the next 6 months, I think it's going to be much easier/cheaper to roll it closer.
7	neutral bearish <Message-ID: > What it's trading what I think it's really worth apr oct 540 500 nov mar 547 375 cal
8	But I guess you know that and that's the point. Again, I think the commercials that showed why we are the most
9	options as the market moves. In terms of straddle strikes, I think the edge received from buying straddles struck on
10	and vice-versa is not big enough to compensate for what I think the industry will view as a scam and another way
11	I agreeed. waiting on his response. <Message-ID: > Jean: I think the location i talked about before is actually better
12	days i'd be worried about a short covering rally. however, i think the market is going to have a hard time placing the
13	orders: 1: When a customer opens up the limit order box, I think the time open should default to 12 hours. We want
14	fyi, they're willing to take us for 20 years. <Message-ID: > i think the velocity of the down move will be much less
15	margin person, John Jones, at 212 469 6773. I think they are trying to margin us. John <Message-ID: >
16	> i loved them in that i thought they were going to go up...i think they still might 3 weeks ago now short term very
17	<Message-ID: > no thx <Message-ID: > i think they worked this out. John <Message-ID: > why is
18	to act weird around each other going forward. Something I think we both wanted to try and we did. Maybe best left
19	you as a good friend. Nothing has changed that nor do I think we need to act weird around each other going
20	be able to offer adequate size on the options. Optimally, I think we want to offer a minimum size of 100 across all

Despite being a Vice President of Enron, Arnold's uses of *I think* are restricted to expressing opinions or tentative predictions (Example 22)—either about the weather (line 2), the industry (line 10), the market (line 12) or other people (lines 15–17)—rather than being used as a way of modifying or mitigating directives.

Example 22

```
<Date: Wed, 9 May 2001 20:04:00 -0700 (PDT)>
<From: john.arnold@enron.com>
<To: jennifer.fraser@enron.com>
<Subject: Re: reminder -pira dinner sund may 13th 7.45 pm st regis>

i think the velocity of the down move will be much less severe from
here.  still dont think this is equilibrium.  need to see aga
coming in lower than expectations for a couple weeks signaling that
we've moved down the demand curve.
```

In contrast, Haedicke's data contains a number of distinctive *I think* sequences, many of which he uses to express indirect directives. Haedicke expresses opinions using the recurring sequence *I think it is a* five times, followed by either *good* or *great* as adjectives (Example 23). Each of the five times he uses this five-gram, he is expressing his personal evaluation of experiences, events or ideas, or in Fetzer's (2014: 68) terms his 'attitude' towards them. Furthermore, the sequence *I think it is a* is only found twelve times in the rest of the Enron corpus, and only one other author, Kay Mann, uses it more than once. When a sixth word *good* is added, as in

Example 23, it becomes entirely unique for Haedicke in the Enron corpus, and as such is distinctive of his idiolect at population level. He also expresses positive opinions and appraisals of his recipients' work through the use of *I think your* (Example 24). In this particular example he is responding to one individual (George McClellan) and has copied three other addressees into the email, and uses *I think* to express two opinions in quick succession. Though not as rare as *I think it is a* in the Enron corpus, *I think your* is fairly distinctive of Haedicke, being used only 16 times by ten different authors in the corpus, only four of whom use it more than once.

Example 23

<From: mark.haedicke@enron.com>
<To: louise.kitchen@enron.com>
<Subject: Re: CACs Forms>

Louise:
Is Enron ready for this? **I think it is a good** suggestion for the most significant deals. It is really positive in that it would force dealmakers to pay more attention to what the contracts provide. [...]

Example 24

<From: mark.haedicke@enron.com>
<To: george.mcclellan@enron.com>
<Subject: Re: Mission UK>
<Cc: richard.sanders@enron.com, robert.quick@enron.com, stuart.staley@enron.com>

George, **I think your** proposal is a good one in light of all the facts at hand. **I think it is** best if we take the high road in approach as you have proposed. [...]

Most recurrent, however, is Haedicke's use of *I think* to reduce the face-threat of directives given to recipients. In Example 25, he is giving instructions to his recipient and colleague Jeff Hodge, directing him to 'start working on a short memo'. This is followed immediately by the sequence 'I think we need to have at least two dates'. As Zhang (2014: 242) notes, *I think* as a mitigator can be 'used to maintain [the] face of the hearer, showing respect and friendliness to someone in [a] lower rank'. Indeed, it might be here that Haedicke uses *I think* to down-tone the effect of having two directives straight after one another. In addition, he uses *we* in such a way that 'softens' the directive, as it is more implicit than using *you* (Vine 2004: 97), for example if he had asserted 'you need to include at least two dates'.

Example 25

<From: mark.haedicke@enron.com>
<To: jeffrey.hodge@enron.com>
<Subject: Re: Antitrust Briefing>

Let's do it in August during the week of the seventh. Please start working on a short memo for Dave and me to send out. **I think we need** to have at least two dates. The memo should say attendance is mandatory.
Mark

As well as these implicit forms, however, Haedicke does use the more explicit *I think you should* (Example 26). This example offers a stark contrast to that in Example 25. Here, Haedicke is writing to his assistant Jane Elbertson, and although the directive is softened by the use of *I think*, he does use the more face-threatening *you* and the deontic modal verb *should* reflecting her obligation to comply with his directive. The relationship between Haedicke and his assistant is likely to be one in which directives such as this are made frequently, and are expected, and so less politeness work is required. This is supported by the fact that the *I think you should* is preceded by the very telegraphic and blunt *please handle.*

Example 26

<From: mark.haedicke@enron.com>
<To: janette.elbertson@enron.com>
<Subject: U.S. Small Business Admin. Evaluation of Enron's Subcontracting Program>

Please handle. **I think you should** send an email to the OGC and ask if they are aware of any such contracts. Mark

In terms of distinctiveness, *I think we need* and *I think we should* are very common in the Enron Corpus, being used by 29 and 46 additional authors respectively. The less implicit *I think you should* is less popular in the Enron corpus, being used more than once by only five other authors. The only verb which Haedicke selects with *I think you should* more than once is *send*, as in Example 26. Once this fifth word is added, no other authors in the Enron corpus use this n-gram more than once, and as such this sequence can be considered to be distinctive of Haedicke at population level.

Finally, Steffes is distinctive from both Arnold and Haedicke in that 44 of his 64 (68.75%) instances of *I think* are followed by a post-predicate *that*-clause (Figure 19), which neither Arnold nor Haedicke use.

Figure 19. 25 of 64 concordance lines for *I think* in Steffes' data

N	Concordance
1	in NE effectively makes NG into a "market participant" - I think that Enron should make as a hold firm position
2	Thanks, Jim <Message-ID: > Agreement - IURC C... I think that Enron is interested supporting your idea.
3	NG application to be managing member of the ARTO. I think that it would be good to call Buckman and ask if
4	Admin will now have the response as its top priority. I think that it will take a couple of weeks before we really
5	activities. I am trying to get a copy of the contract, but I think that it is in place until 8/02. My recommendation
6	language that counters the below language - I think that the state-level parties (and the IURC) have an
7	next year, probably after SCE gets its money. Finally, I think that the right policy decision is not to "re-bill"
8	put in an RCR to discuss ASAP (probably next week). I think that the key issue involves the inclusion of
9	markets work. Finally, on the issue of blanket authority, I think that the CAISO model is a good example of why
10	-- Here is the org chart that captures our discussions. I think that this org and the # of people is the lowest level
11	Mechanism and for Order Shortening Time Harry -- I think that this makes alot of sense. I like working
12	etal to discuss next Tuesday? Jim <Message-ID: > I think that this list is fine. Any thoughts to including the
13	My thoughts -- 1. Focus more on Tariff standardization (I think that this is the key issue for Enron because LMP
14	rate adjustment to re-capture their previous mistake). I think that this is something we need to go into the
15	resolve the question of the Native Load Exception. I think that this is critical to highlight if this issued is
16	Group. The fiefdom grows!!!! Jim <Message-ID: > Ray -- I think that this is about the CalPX bankruptcy. Any
17	a strong preference for what DSTAR utilities should do? I think that we want to communicate with OMTR and
18	I take the "delay" side for $200? Jim <Message-ID: > I think that we should support the effort. I would not get
19	sure that you had seen the note Sanders sent around? I think that we should discuss today on the 4:30 call. Jim
20	did we want to do with the confidentiality of our credit (I think that we should agree - not PJM's to distribute).
21	the discussions at NERC recently to broaden their role, I think that we need to be very coordinated on our EISB
22	some elements that NY thought was better than PJM. I think that we need to "find" three things to give to the
23	soon as possible. Thanks, Jim <Message-ID: > Harry -- I think that we need to talk with EWS Tariff risk to make
24	> Lisa -- Rick Shapiro has offered to present. I think that we can work up some presentation for 45
25	want EPSA to focus on? If you send me your thoughts, I think that we can send to EPSA before the phone call.

The trigram *I think that* is infrequent in the Enron corpus; of all 2,815 instances of *I think* in the corpus, only 375 (13.32%) are followed by *that*, compared with 68.75% in Steffes' data. As such, the choice to retain *that* in post-predicate clauses is distinctive of Steffes when his usage is compared with the Base Rate Knowledge. By far the most preferred choice is the zero-*that* variant, as exhibited by both Arnold and Haedicke. Because the choice to include the relativizer *that* is an unusual one in the Enron population, subsequent extensions of this n-gram in Steffes' emails are also rare. For example, in comparison with Arnold who uses *I think the* to express opinions, Steffes uses *I think that the* four times (Example 27).

Example 27

<From: jim.steffes@enron.com>
<To: susan.lindberg@enron.com>
<Subject: RE: EPSA conference call re. Reporting Requirements NOPR>

Susan --
Please put in an RCR to discuss ASAP (probably next week).
I think that the key issue involves the inclusion of financial transactions. Need a solid message from industry that this is inappropriate - which EPSA members are filing this info?
Jim

In the wider Enron corpus, *I think the* is found 204 times, compared with *I think that the* which is only found 35 times, four of which belong to Steffes. Besides Steffes, only seven of the 175 additional authors use *I think that the* more than once, compared with 35 authors for *I think the.* So, when *the* is the initial word in the subordinate clause after *think*, the plus *that* variant is far rarer. Similarly, in contrast to Haedicke who uses *I think we should* to mitigate directives, Steffes uses *I think that we should* and *I think that we need to* in such a way that reduces the face-threat of the directive he is giving to his recipient. In Example 28 and 29 Steffes is writing to multiple addressees, two of whom are addressees in both emails (sarah.novosel@enron.com, l.nicolay@enron.com). In both, Steffes marks topics and asks questions, suggesting that he is the superior, or one of the superior participants in these email conversations. In addition, he makes suggestions for action for the group, which he mitigates with *I think that*. When tested against the rest of the Enron corpus, *I think that we should* is used a total of 14 times and only two other authors use it more than once. Furthermore, *I think that we need to* appears only four times in the rest of the corpus, all of which are used by different authors, meaning that Steffes is the only author in the Enron corpus to use this lexical sequence more than once.

Example 28

<From: jim.steffes@enron.com>
<To: sarah.novosel@enron.com,
<l.nicolay@enron.com,steve.walton@enron.com,>
<paul.kaufman@enron.com>
<Subject: RE: I'm Back on Email>

Good news. Also, Rich Drom called from PJM wanting to know if someone from Enron was attending the Members Mtg this Thursday (no) and if what did we want to do with the confidentiality of our credit **(I think that we should** agree - not PJM's to distribute).
Your thoughts - I'll call Drom back tomorrow.
Jim

Example 29

<From: jim.steffes@enron.com>
<To: sarah.novosel@enron.com, richard.shapiro@enron.com,>
<steve.montovano@enron.com, tom.hoatson@enron.com,>
<l.nicolay@enron.com, howard.fromer@enron.com,>
<daniel.allegretti@enron.com>
<Subject: RE: Lunch with Gail McDonald (MD PSC)>
<Cc: linda.robertson@enron.com>

Great news. Are there any other MD Commissioners that Ms. McDonald thinks we should visit with?
On the idea of NYISO vs. PJM, I was reading in Restructuring Today yesterday about some elements that NY thought was better than PJM. **I think that we need** to 'find' three things to give to the NYISO in the development of a single market...allocation of FTRs comes to mind...

Stylometric studies into authorship analysis have traditionally focused on very high frequency words in their categorisation of documents. However, focusing on frequency alone is not enough in order to make claims about individuals' idiolects. This analysis of *I* has found that, even with the most high frequency words, authors' collocational patterns and preferences can quickly become distinctive. *I* is the third most common word in the Enron corpus, and it occurs significantly more frequently than in American English generally, as represented by *COCA*. In turn, different employees within the corpus use *I* with different frequencies, with some (Dorland, Germany and Lavorato) using it significantly more frequently than is expected in the Enron population.

Analysis of the most frequent R1 collocates of *I* in the corpus offers a Base Rate Knowledge as to the words which are most commonly co-selected with it. Of these collocates, *will* and *think* were third and fourth most common respectively, and present interesting examples for study, as they both perform pragmatic and epistemic functions in the workplace. Again, some authors use these collocates more than others. For example, Derrick and Nemec were identified as high-frequency *I will* users. A subsequent qualitative analysis of how they use *I will* found that these two authors use this collocation in different ways. While Derrick uses it most often to courteously reject event invitations and offer promises of support to his recipients, Nemec's uses are far more influenced by his collaborative engagement in the drafting of legal documents. These different functions of use produce linguistic output from these two authors that not only distinguishes them from one another, but also is very rare, if not non-existent, in the rest of the Enron Email Corpus. At the same time, Kaminski, a non-native English speaker, who was taught English rather

than acquiring it as a first language, overwhelmingly prefers *I shall* over *I will*, a bigram only he uses in the whole Enron corpus. Similar results were found with *I think*, a feature with various functions in the workplace. On the basis of relative frequencies and a Base Rate Knowledge of how often *I* was co-selected with *think*, three authors were identified as significantly high-frequency users: Arnold, Haedicke and Steffes. While Arnold's data did not exhibit any distinctive word n-grams, the function with which he used *I think* differed from that of the other two authors. He reserved his use of it for expressing opinions and hedging predictions. On the other hand, Haedicke and Steffes also used *I think* to express opinions, but also to mitigate directives. Despite using *I think* for similar purposes, both of these authors produced word n-grams which were distinctive at population level when tested against the rest of the corpus. Most notable was Steffes' choice to retain post-predicate *that*, which is uncommon in the Enron corpus.

These findings, then, show that even when the focus is on very frequent words (in this case a function word) or collocations, authors produce lexical sequences and word n-grams that can be considered distinctive. They are distinctive either insofar as they are used by a very small number of other authors, or that the author in question is the only one in the Enron population who uses them at all. When such population-level distinctiveness is identified, a case can be made that these lexical strings are observable manifestations of authors' idiolects. In turn, this goes some way towards providing empirical evidence to substantiate the theory.

5.3 Content words and idiolect: the case of *deal*

As underlined by the review of stylometric research into authorship analysis (Section 2.2.2) content words have largely been avoided as features used to distinguish between and identify authors. Despite the promising results of studies which have drawn on content words *as well as* function words (e.g. Hoover 2003; Diederich et al. 2003; Labbé 2007; Jockers and Witten 2010), there is still the belief that 'function words are better indicators of authorship than content words' (Grieve 2007: 261). This is largely due to the fact that content words are too heavily related to the topic of the text in question, and so any similarity between texts or authors on the basis on content words is at risk of being because they share a topic rather than a style (Holmes 1994: 90; Koppel et al. 2009: 11–12). However, Argamon and Koppel (2010) emphasise the value of content words in automated text classification, noting

that ‘at times textual features that mainly indicate topic (e.g. “content words”) may be useful for a form of stylistic discrimination’ (Argamon and Koppel 2010: 81). In their model of the ‘communicative act’ and the contextual elements than influence the construction of any text, they incorporate ‘content’, along with ‘author’, ‘audience’, ‘medium’ and ‘purpose’, commenting that:

> Certainly there are correlations between these different facets—particular authors will have idiosyncratic topical preferences, certain topics are largely embedded in communal discourses that come with particular stylistic commitments…
>
> (Argamon and Koppel 2010: 84)

In Coulthard’s (2013: 448) report of the case involving company emails, he found that the most distinctive linguistic features in the disputed email were content words and the co-selection of these content words within the same text. A number of these were related to the company context in which the text was produced, words such as ‘employees’, ‘competitors’, ‘fully expensed’, and ‘recognised revenue’. When comparing the disputed email with the known documents of the potential suspects, these lexical choices were found to be consistent with those made by one of the authors in his known documents. Specifically, he was an accountant who wrote frequently elsewhere about ‘recognising revenue’ and used the phrase ‘fully expensed’ (Coulthard 2013: 457). What further strengthened the evidence was that these phrases were not used anywhere else by the other suspects, or by anyone else in Coulthard’s 190,000 email database. In this case, distinctive use of content words, and in particular register-related content words, were useful in determining authorship.

The analysis in this section shows that by using a corpus linguistic approach, and relying on the constructs of Base Rate Knowledge and population-level distinctiveness, content words can be central to the investigation of idiolect. The specific word which is the focus of this analysis is *deal*. This word is ‘embedded in communal discourses’ (Argamon and Koppel 2010: 84) of Enron, as it relates to the core business of Enron as an energy trading company. *Deal* in this context, refers to the verb senses: ‘to distribute or bestow among a number of recipients’ and ‘to carry on commercial transactions; to do business, trade, traffic’, and the noun sense: ‘an act of dealing or buying and selling; a business transaction’ (*Oxford English Dictionary*). The importance of this word in Enron is attested to by its frequency.

Deal occurs 4,134 times in the Enron corpus, accounting for 0.17% of the total 2,462,151 tokens, and is used by 125 of the 176 of the authors. This proportion represents the Base Rate Knowledge of how frequently this word is used in the Enron population. In comparison, *deal* occurs 87,551 times in *COCA*, accounting for only 0.02% of the 464,020,256 tokens.

Some authors use *deal* more frequently than others in the corpus. Of the EEC12 authors, Germany and Farmer are the two most frequent users, and the only two authors who use it significantly more frequently than the Base Rate 0.17% (Table 12).

Table 12. Frequency of *deal* across authors in EEC12

Author	**Tokens**	***Deal* (base rate =0.17%)**
Farmer	**24,502**	**286 (1.17)***
Germany	**77,597**	**693 (0.89)***
Allen	14,579	4 (0.03)
Zipper	7,393	13 (0.18)
Lavorato	24,677	41 (0.17)
Nemec	50,211	71 (0.14)
Arnold	26,283	15 (0.06)
Dorland	14,565	12 (0.08)
Steffes	35,284	25 (0.07)
Haedicke	18,540	19 (0.10)
Kaminski	52,992	4 (0.01)
Derrick	5,902	0 (0.00)

As a result, it is their use of *deal* that is analysed here. Beyond the EEC12 sample, these two authors rank highly in their use of these words within the full Enron corpus. In terms of relative frequency of *deal* as a proportion of all their tokens, Farmer ranks as third highest ranked of 176 authors and Germany ranks as sixth. However, the top two users have fewer than 600 tokens in their dataset and fewer than ten hits for *deal.* With these authors removed, Farmer becomes the most frequent *deal* user, with Germany in fourth place overall in the Enron Email Corpus.

The criticisms stylometrists target at content words are that they are indicative of topic rather than idiolect. This criticism, it could be argued, is justified when such quantitative studies rely only on the relative frequencies of individual words. The occurrence of *deal* in a hypothetical disputed document and the known documents of one of the authors (e.g. Farmer or Germany) is not reliable evidence that the texts are written by the same person. As Coulthard (2013: 447–8) points out, however, it is the co-selection and collocations chosen in relation to these topic-sensitive words that are more important for authorship attribution. The analysis here, therefore, sets out to identify the different ways in which Germany and Farmer use *deal*, and how such differences can not only distinguish them from each other, but also at population level. First, the authors' 'collocational profiles' for this word are compared against each other and against the Base Rate Knowledge for the corpus. This is followed by an identification of all the word n-grams Germany and Farmer use which include *deal*, and an investigation as to how many of these are unique to these two authors. Overall, the findings show how a corpus linguistic approach to analysing linguistic variation can place content words at the centre of discussions about idiolect.

5.3.1 Comparing collocational profiles of *deal*

The term 'collocational profile' is used by corpus linguists to refer to the words occurring in the immediate environment of the node word (e.g. Sinclair 1996; 2004, Tognini-Bonelli 2001; Gilquin 2010; Kehoe and Gee 2009; Manca 2010). In this analysis, because the focus is on recurring lexical sequences and word n-grams, the collocational profile of *deal* is taken to be everything within five words to the right and left of this word (L5–R5 collocates). This kind of profile is also known as the 'collocational horizon' (Scott 2008b) and the collocational 'span' (Stubbs 2001: 29) of a word. Figure 20 shows the collocational profile of *deal* in the Enron Email Corpus computed by *Wordsmith Tools* (see Section 4.1.2). It shows the collocates organised in terms of frequency within each column, so *the* is the most common L1 collocate, while *with* is the most common R1 collocate, and so on. This collocational profile can serve as the Base Rate Knowledge in terms of the collocational environment(s) in which *deal* most frequently finds itself in the Enron corpus, and how it is normally used by Enron employees in their emails.

Figure 20. The collocational profile of *deal* in the Enron Email Corpus

N	L5	L4	L3	L2	L1	Centre	R1	R2	R3	R4	R5
1	THE	THE	THE	THE	THE	DEAL	WITH	THE	THE	THE	THE
2	ID	TO	TO	A	THIS		IS	MESSAGE	ID	TO	TO
3	TO	ID	A	OF	A		AND	IS	MESSAGE	ID	MESSAGE
4	MESSAGE	MESSAGE	ID	ON	TO		I	FOR	TO	AND	ID
5	I	A	IS	TO	ON		WAS	I	AND	I	A
6	IS	I	I	ID	NEW		TO	TO	IS	MESSAGE	I
7	A	IS	MESSAGE	FOR	BIG		FOR	AND	I	A	AND
8	AND	THIS	FOR	IS	ID		IN	ID	A	IS	WITH
9	FOR	AND	ON	HAVE	CREATED		NUMBER	IN	FOR	DEAL	YOU
10	THIS	ON	OF	IN	IN		MESSAGE	WITH	IN	THIS	IS
11	WE	FOR	AND	AND	OF		NUMBERS	A	ON	IN	DEAL
12	DEAL	DEAL	IN	I	AND		TICKET	YOU	CHANGED	FOR	THIS
13	IN	YOU	YOU	MESSAGE	FOR		HAS	THIS	THIS	THAT	THAT
14	YOU	OF	NOT	THAT	THAT		THAT	BEEN	WE	WE	FOR
15	OF	WE	THIS	AN	GREAT		ENTRY	ARE	BE	OF	IT

The profile indicates, for example, that *deal* is very commonly used as a noun in *the deal* (which occurs 790 times) and a verb in *deal with* (which occurs 273 times). Longer recurring sequences include *deal with the* and *deal has been*, which occur 53 and 50 times respectively. What this profile does not indicate is how frequently *deal* is used as a noun and is followed immediately by an identification number, when Enron employees are emailing about specific deals, such as in:

Example 30

<From: kate.symes@enron.com>
<To: evelyn.metoyer@enron.com>
<Subject: Re: 11-13-00 Discrepancies>

Deal 456717 has been changed to reflect Pinnacle West Capital Corp, and the traders have been reminded of the name change. [...]

This pattern is not represented in the collocational profile given that the same number is rarely repeated as an R1 collocate; the most commonly occurring number is *deal 895826* which occurs five times. However, the overall pattern *deal + #* occurs 657 times, making numbers (as a collective group) the most common, or conventional, R1 collocates of *deal* in Enron. In addition, *number* and *numbers* both appear as frequent R1 collocates of *deal* in Figure 20, occurring a total of 159 times. On 20 occasions, *deal number(s)* is followed directly by a number, such as:

Example 31

<From: jane.tholt@enron.com>
<To: laurie.ellis@enron.com>
<Subject: Re: San Diego - Deal 341242>

I switched **deal number 341242** to a sale and deal number 341270 as a buy

This Base Rate collocational profile presented by Figure 20 can be compared with the collocation profiles of *deal* in Germany's (Figure 21) and Farmer's emails (Figure 22).

Figure 21. Collocational profile of *deal* in Germany's data

N	L5	L4	L3	L2	L1	Centre	R1	R2	R3	R4	R5
1	THE	THE	THE	THE	THIS	DEAL	TICKET	MESSAGE	ID	THE	TO
2	I	MESSAGE	ID	ON	THE		IS	FOR	THE	AND	DEAL
3	DEAL	I	I	I	ON		WITH	AND	TO	I	I
4	TO	ON	TO	A	CREATED		NUMBER	WITH	MESSAGE	TO	THE
5	ID	DEAL	A	FROM	WITH		TICKETS	I	IS	ID	ID
6	WE	FOR	FOR	TO	TO		I	IS	A	MESSAGE	MESSAGE
7	FOR	A	ON	OF	A		TO	TO	I	DEAL	THIS
8	AND	TO	DEAL	FOLLOWING	CES		NUMBERS	THE	DEAL	THIS	OF
9	OF	AND	YOU	BOOKOUT	AND		IN	FROM	THIS	A	A
10	IS	AT	GAS	PLEASE	SITARA		FOR	VOLUME	BE	WE	AT
11	MESSAGE	ID	MESSAGE	AND	NEW		MESSAGE	THIS	1	IT	AND
12	THIS	OF	IT	IN	PURCHASE		AND	SHOULD	AND	NEW	FOR
13	CES	CHANGED	THIS	FOR	DEALS		ON	WE	SALE	1	ON
14	PLEASE	DEALS	OF	ID	CITYGATE		FROM	ID	WE	IS	YOU
15	AT	IS	TERM	VOLUME	FOR		THE	IN	SOLD	FOR	WITH

Figure 22. Collocational profile of *deal* in Farmer's data

N	L5	L4	L3	L2	L1	Centre	R1	R2	R3	R4	R5
1	MESSAGE	ID	I	THE	THE	DEAL	TICKET	TO	THE	ID	MESSAGE
2	THE	MESSAGE	ID	HAVE	THIS		TO	FOR	MESSAGE	THE	THE
3	D	THE	THE	A	CREATED		FOR	D	COVER	D	D
4	ID	D	MESSAGE	I	TO		WITH	IS	ID	THIS	TO
5	IS	I	TO	ID	NEW		D	THE	THIS	FOR	I
6	A	WE	D	TO	ON		IN	MESSAGE	SHOULD	TO	OF
7	I	TO	ON	ON	ID		HAS	HAVE	IS	IN	YOU
8	DEAL	YOU	CAN	MESSAGE	SPOT		IS	YOU	FOR	2	IN
9	FOR	A	OF	FOR	A		I	IN	SITARA	I	DEAL
10	WE	THIS	HAVE	OF	ROLLED		AND	SHOULD	DEAL	MESSAGE	MTR
11	TO	AND	IS	CORRECTED	EXTENDED		AT	THIS	ZERO	BE	AS
12	HAVE	CHANGED	DEAL	PRICING	TERM		NUMBERS	UNDER	IN	VOLUME	ON
13	YOU	RECORD	THIS	CHANGE	ADJUSTED		THE	BEEN	AND	ARE	FLOW
14	DAREN	SHOULD	AND	ALLOCATED	NO		WAS	HAS	VALID	GAS	METER
15	CAN	FROM	PRICE	EXTENDED	SWING		LET	NOT	BEEN	IF	DEALS

Through these comparisons, we can see how Germany's and Farmer's use of *deal* contrasts with each other and stands out against the corpus as a whole. The red highlighted words in the Figures are those words that are not part of the *deal* collocational profile for either the other author or the Enron corpus generally. Again, what is not shown in these profiles is that Germany and Farmer are both prolific users of the *deal* + # pattern. This occurs 362 times in Germany's emails, accounting for just over half (52.24%) of his 693 instances of *deal,* and for 110 (38.46%) of Farmer's 286 uses of *deal*. By comparing these collocational profiles, the ways in which Germany and Farmer are similar to, and different from, the Base Rate Knowledge of the corpus are clear. There are similarities across the profiles, especially in the most frequent collocates such as *the*, *this* and *created* in L1 position, and *ticket*, *with* and *is* in R1 position. Despite this, the abundance of

highlighted words in the profiles for Germany and Farmer shows that both have distinctive collocational profiles for *deal*, not only when compared with each other, but when compared with how *deal* is normally used in the population. In turn, these highlighted words can provide a point of access through which author-distinctive co-selection patterns and lexical sequences can be identified. Some of these will be analysed qualitatively here.

First, *with* is the fifth most frequent L1 collocate of *deal* in the collocational profile of Germany but not for Farmer or the Enron corpus as a whole. In fact, the unremarkable collocation *with deal* is only found 22 times in the whole corpus and eighteen of those are from Germany's dataset, all but one of which are followed by a deal number (Figure 23).

Figure 23. 18 of 18 concordance lines for *with deal* in Germany's emails

N	Concordance
1	of 688 day for the 1st - 31st). Bookout deal 533317 with deal 533319 (volume of 5000 day for the 5th - 31st).
2	happy. <Message-ID: > Please bookout deal 563831 with deal 563837. <Message-ID: > Please let me know
3	> Please bookout the following deals; Deal 643754 with deal 759487 for May and June. <Message-ID: > Did
4	> Please bookout the following deals; Deal 643754 with deal 759487 for May and June. <Message-ID: > Did
5	for the month of July. I created deal 315460 to bookout with deal 217769 and deal 315471 to bookout with deal
6	bookout with deal 217769 and deal 315471 to bookout with deal 223967. Please let me know if you have any
7	deals 229344 and 229357. Should it also be a bookout with deal 227882? What about deals 169036and
8	Mon from New Power (deal 525128) at $8.40. Bookout with deal 525121 (sale at IF + .0125). <Message-ID: >
9	597309 Buy Dynegy 8000 IF + .05 Monclova - bookout with deal 593311 593311 Sale New Power 8000 IF + .05
10	Sale New Power 3578 IF + .02 Paulding - Bookout with deal 597295 597309 Buy Dynegy 8000 IF + .05
11	597295 Buy Dynegy 3578 IF + .02 Paulding - Bookout with deal 597302 597302 Sale New Power 3578 IF + .02
12	States deals in Unify on the 31st. I remember chatting with deal clearing about it yesterday. Could you path
13	168996. I took the expected volumes to 0. This goes with deal 643761. <Message-ID: > Have I told you how
14	> Please match deal 203315 (CES sale for Mar) with deal 209122. - <Message-ID: > Please match deal
15	for the 1st - 4th, 35,133 of the transport will be matched with deal 268094 and the balance of the transport will be
16	> Please match deal 204176 (CES sale Mar-May) with deal 209005 (what do you think, CES buy
17	<Message-ID: > Deal 125925 has been replaced with deal 231766 and deal 125928 has been replaced
18	with deal 231766 and deal 125928 has been replaced with deal 234584. Angie is all over this one.

Besides Germany, none of the other 176 authors uses the bigram *with deal* more than once. Therefore, this bigram can be considered as being distinctive of his idiolect; although there are over four thousand instances of *deal* in the corpus, Germany is the only one of 176 authors to collocate it with *with* in L1 position more than once.

Concordance lines 5–11 in Figure 23 also reveal another distinctive element of Germany's collocational profile of *deal*: the use of the verb *bookout* in L2 position. According to investment education website *Investopedia.com*, *bookout* is the act of 'closing out an open position in an OTC derivative, such as a swap

contract, before it matures'; in turn, *OTC* is an abbreviation of 'over-the-counter', which 'can be used to refer to stocks that trade via a dealer network as opposed to on a centralized exchange'. It is clearly a specialised term, yet it is used only 36 times in the corpus, 33 of which are used by Germany, and the other three are used by John Forney, a manager. The trigram *bookout with deal* is used by Germany in emails to his colleagues either instructing them to 'bookout' particular deals, or informing them that he has:

Example 32

<From: chris.germany@enron.com>
<To: alvin.thompson@enron.com, joann.collins@enron.com>
<Subject: Bookout>
<Cc: jeffrey.porter@enron.com>
<Bcc: jeffrey.porter@enron.com>
I just purchased 35000 day for Sat, Sun and Mon from New Power (deal 525128) at $8.40. **Bookout with deal** 525121 (sale at IF + .0125).

Example 33

<From: chris.germany@enron.com>
<To: victoria.versen@enron.com, alvin.thompson@enron.com>
<Subject: CES Deals for July>
<Cc: molly.johnson@enron.com>

I show no sales to CES for the month of July. I created deal 315460 to **bookout with deal** 217769 and deal 315471 to **bookout with deal** 223967. Please let me know if you have any questions.

Germany is the only author in the Enron corpus to use the trigram *bookout with deal*. *Bookout* and, to a lesser extent, *deal* are both specialised terms in Enron. It is this type of situation in which the Enron Email Corpus is valuable as a reference corpus representing a relevant population from the same speech community. If Germany's use of this trigram is unique within against this population, wherein there are many other traders and employees writing about deals, then this is strong evidence to suggest that it is part of his distinctive idiolect. Indeed, a Google search for '*bookout deal with*' returns only five results, all of which are from Germany's emails within online versions of the Enron corpus.

As well as *bookout*, another L2 collocate which features in Germany's *deal* collocational profile and not Farmer's, nor the Enron corpus more widely, is *from*. The 'collocational framework' (Renouf and Sinclair 1991: 129) *from + x + deal* appears a total of 24 times in the Enron corpus, 19 of which are found in Germany's

emails, and he is the only author to use it more than once. Almost all of these instances are found in emails in which Germany is making reference to a particular deal with a particular company, and follows the company name immediately with the deal identification number (Example 34–35).

Example 34

<From: chris.germany@enron.com>
<To: robert.allwein@enron.com, joann.collins@enron.com>
<Subject: Re: Columbia Gas of Ohio on CGAS for Jan 00>
<Cc: dick.jenkins@enron.com, sandra.dial@enron.com>

Since we now feel certain ENA bought this gas on Jan 21st, please path the supply deal **from COH (deal** 153863) to one of those large CES deal tickets.

Example 35

<From: chris.germany@enron.com>
<To: joan.veselack@enron.com, robert.allwein@enron.com>
<Subject: CPA purchase>
<Cc: victor.lamadrid@enron.com>

We bought 4385 **from CPA (deal** 155237) and sold it to CES (deal 155238).
We bought 1249 **from CPA (deal** 155240) and sold it to EES (deal 155244).
Do the emails help?? Is this the way you guys want ot see this?

At the same time, at the other side of the node word, the preposition *from* in R2 position is also a distinctive element of Germany's collocational profile of *deal*. The reverse collocational framework *deal* + *x* + *from* is found 31 times in the Enron corpus, and 20 of these belong to Germany. Unlike the first pattern, though, two other Enron employees use *deal* + *x* + *from* more than once (Kate Symes and Dutch Quigley). In Germany's emails, the *x* element in *deal* + *x* + *from* is invariably filled with a deal identification number. In using this trigram, Germany is either notifying his recipient that he has made a change to some details regarding a particular deal (Example 36), or is instructing them to do so (Example 37).

Example 36

<From: chris.germany@enron.com>
<To: crystal.hyde@enron.com, kyle.lilly@enron.com>
<Subject: Transport Usage>

The exchange deal numbers are 323558 and 323553.
Also, I changed the rate on Tenn **deal 235293 from** $.11 to $.0097.
thanks

Example 37

<From: chris.germany@enron.com>
<To: kimat.singla@enron.com>
<Subject: Deal 806589>

Please change the counterparty on **deal 806589 from** TP2 to TP3
(sorry about that).

Despite *deal* being used so frequently by the Enron population, these two collocational patterns are either entirely distinctive of Germany (in the case of *from + x + deal*), or very rarely used elsewhere (in the case of *deal + x + from*). In terms of idiolect, then, although *deal* is a word which is very common in the population from which he is taken, the way in which he encodes information about deals is very distinctive.

Shifting focus to Farmer, the most notable difference in his collocational profile (Figure 22) is the proliferation of verbs in L1 and L2 positions that are present in neither Germany's profile nor that of the Enron Corpus generally. Most of Germany's distinctive L1 collocates highlighted in red in his collocational profile of *deal* (Figure 21) are premodifying *deal,* such as *CES* (a company name), *Citygate* (a station at which a gas distributor gets gas from a natural gas pipeline company) or *purchase* (a type of deal). In contrast, many of Farmer's L1 and L2 collocates are verbs, such as *rolled deal*, *extended deal* and *adjusted deal*, as well those which include the definite article, such as *corrected the deal*, *changed the deal* and *extended the deal* (Figure 24).

Figure 24. 25 of 286 concordance lines for *deal* in Farmer's data

N	Concordance
1	Corp. <Message-ID: > I have adjusted deal 452491 for mtr 20014903. We agree
2	D <Message-ID: > I have adjusted deal #152638 to cover the first 20 days of
3	<Message-ID: > Megan, I adjusted deal 529856 on the 2nd of Oct and Nov
4	Corp. <Message-ID: > I have extended deal 151669 to cover mtr 9676 for Feb
5	Corp. <Message-ID: > I extended deal 559483 through Dec 01. D
6	the past? D <Message-ID: > I extended deal 93481 for another year. D
7	copied. D <Message-ID: > I extended deal 583232 to cover 2/1 and #604056 to
8	Amiga. D <Message-ID: > I extended deal 461059 for the rest of October. D
9	Rivers? D <Message-ID: > I have rolled deal 506192 thru 1/1/01, priced at Dec
10	be cleared. D <Message-ID: > I rolled deal 331917. <Message-ID: > Done.
11	in place. Daren <Message-ID: > I rolled deal 418382 for Nov 1st. D <Message-ID:
12	and 16th. D <Message-ID: > I rolled deal 150325 for the first 3 days of Jan. I
13	this meter? <Message-ID: > I rolled deal 128952. <Message-ID: > I have
14	questions. D <Message-ID: > Rolled deal 454057 to cover flow at mtr 5192. d
15	The correct price is 4.50. I changed the deal ticket. D <Message-ID: > I'll be
16	fine <Message-ID: > I've changed the deal to 1 for the rest of this month and
17	is the correct price. I have changed the deal ticket. D <Message-ID: > Thanks!
18	> Done. <Message-ID: > I corrected the deal. The price should be Waha Index
19	should be HSC GD-.04. I corrected the deal ticket. D <Message-ID: > Danny,
20	price is 5.18. I have corrected the deal ticket. D <Message-ID: > I have
21	final agreement. I have corrected the deal in Sitara. D <Message-ID: > KH,
22	d <Message-ID: > I've extended the deal for the rest of the year, with a -0-
23	only. D <Message-ID: > I extended the deal. D <Message-ID: > Apply this
24	D <Message-ID: > I extended the deal at mtr 6719 to include 2/10. In the

As shown in the Figure, the subject of all of these verbs is *I*, and as such the adjusting, extending and rolling of deals appears to be part of Farmer's role in the corporation. His uses of these collocations are all found in emails in which he is informing his recipient that he has acted (Example 38–40).

Example 38

```
<From: daren.farmer@enron.com>
<To: jackie.young@enron.com>
<Subject: Re: Meter # 0986725 - 1/00 production - Encina Gas
Marketing>

I have adjusted deal #152638 to cover the first 20 days of flow in
Jan.
```

Example 39

<From: daren.farmer@enron.com>
<To: aimee.lannou@enron.com>
<Subject: Re: Jan '01>
<Cc: edward.terry@enron.com>

I rolled deal 150325 for the first 3 days of Jan. I expect this point to be zero for the rest of Jan.

Example 40

<From: daren.farmer@enron.com>
<To: mary.poorman@enron.com>
<Subject: Re: Meter 6315, purch from Torch/Rally, October>

I extended deal 461059 for the rest of October

Many of these verbs, as shown in the examples, relate to changing the timing of deals, and this may give an indication as to his responsibilities for deals in the company. In comparison with Germany who books out, buys, sells and changes the rate of deals, Farmer appears to adjust their duration or correct their ticket. In terms of distinctiveness, Farmer is the only author in the Enron corpus to use the bigrams *adjusted deal* and *rolled deal*, and the only author to use *extended deal* more than once. Similarly, the trigram *corrected the deal* is only found in Farmer's emails, and he is the only employee to write *extended the deal* more than once. These lexical sequences, then, are distinctive of Farmer at population level, and potentially provide strong evidence of idiolect. Furthermore, a Google search for *I adjusted deal*, returns only three results, one of which is a version of one of his emails. At the same time, *I rolled deal* returned six results that were not emails written by Farmer, *I extended deal* and *I corrected the deal* both returned eight results. Therefore, from the billions of documents searched for by Google, these lexical sequences, found to be distinctive of Farmer in the Enron population, are also very distinctive of him beyond the Enron population.

Most of Farmer's distinctive sequences are used in the *I* + V-ed + *(the) deal* framework. This pattern occurs a total of 71 times in the Enron corpus and 58 of these are found in the emails of Germany (32) and Farmer (26). Although this is a grammatical pattern that Farmer shares with Germany, their verb choices differ. In comparison with those used by Farmer, Germany uses a different range of verbs including *created deal* (14) and *killed deal* (5). While this particular pattern is shared between these two authors, the one in which the auxiliary verb *have* is inserted

before the main verb is entirely distinctive of Farmer. It appears a total of 21 times in the Enron corpus, of which he is responsible for twenty. Therefore, it is not only the precise sequence of lexical items around *deal* that is distinctive of Farmer here, but the collocational framework also appears to be a unique element of his idiolect.

This analysis of Germany and Farmer's collocational profile of *deal* has identified how their use of these words stands out against the Base Rate Knowledge of the population from which they are taken. A number of distinctive elements of their collocational profiles have been analysed qualitatively here and potentially strong idiolectal evidence has been revealed in the process. This goes some way towards demonstrating the value of content words— in particular how these content words are used by authors—in the identification of idiolect. However, the qualitative analysis here has been necessarily selective, and the idiolectal nature of content word use is far more pervasive in the data for these authors than has been discussed here. As such, a complete overview of Germany's and Farmer's idiolectal use of *deal* is offered in the following section.

5.3.2 Germany's and Farmer's distinctive *deal* n-grams

A corpus linguistic approach can be used to highlight the exact extent to which *deal* is useful in revealing idiolectal linguistic variation in the emails of Germany and Farmer. Using *Wordsmith Tools* (Section 2.2.2) all of the two to six word n-grams in Germany's and Farmer's data that contain *deal* and are used twice or more can be identified. The limit of six word clusters is based on the collocational profiles of the words analysed above, which ranged from one to five collocates from the node to both the left and right. In total across all n-gram lengths there are 276 *deal* n-grams in Germany's emails and 132 in Farmer's. A remarkable number of these are distinctive of the two authors. Distinctive here means that either Germany or Farmer is the only one of the 176 authors in the Enron corpus to use them more than once. Germany has 108 such distinctive *deal* n-grams, while Farmer has 32 (Table 13). For Germany many of these relate to specific deals, such as *CES deal*, *CGAS deal*, *from CPA deal, from COH deal,* and *bookout and include in citygate deal.* In contrast, none of Farmer's distinctive n-grams refers to specific deals.

Table 13. Germany's and Farmer's distinctive *deal* n-grams

		Germany
Length	**N**	**Examples (frequency)**
2	31	*with deal* (18), *ces deal* (16), *citygate deal* (10), *supply deal* (9), *sales deal* (8), *loan deal* (8), *deal comment* (7), *bookout deal* (7), *singer deal* (6), *match deal* (6), *COH deal* (6), *at deal* (5), *capacity deal* (5), *CPA deal* (5), *market deal* (4), *management deal* (4), *CGAS deal* (4), *deal market* (4)
3	48	*bookout with deal* (10), *volume on deal* (9), *to CES deal* (7), *park loan deal* (6), *the purchase deal* (5), *from CPA deal* (5), *from COH deal* (5), *I killed deal* (5), *please match deal* (5)e, *and created deal* (5), *price on deal* (4), *supply deal market* (4), *deal ticket and* (4), *does this deal* (4), *just created deal* (4), *just killed deal* (4)
4	19	*supply deal market deal* (4), *it to CES deal* (4), *to bookout with deal* (4), *the volume on deal* (4), *the price on deal* (4), *the term on deal* (3), *this deal will be* (3), *the following deal tickets* (3), *a look at deal* (3), *I just created deal* (3), *I just killed deal* (3).
5	7	*sold it to ces deal* (4), *the deal in the system* (4), *take a look at deal* (3), *the volume on this deal* (3), *changed price on deal* (3), *include in citygate deal* (3), *on this deal to zero* (3)
6	3	*for the term of the deal* (4), *and sold it to CES deal* (4), *bookout and include in citygate deal* (3)
Total	**108**	
		Farmer
Length	**N**	**Examples (frequency)**
2	5	*spot deal* (8), *rolled deal* (6), *extended deal* (5), *deal allocate* (3), *adjusted deal* (3)
3	14	*have created deal* (14), *a spot deal* (7), *the term deal* (5), *corrected the deal* (4), *I extended deal* (4), *I rolled deal* (4), *deal however the* (3), *deal to the* (3), *deal numbers you* (3), *allocated to deal* (3), *pricing on deal* (3), *extended the deal* (3),
4	9	*I have created deal* (14), *of a spot deal* (4), *the pricing on deal* (3), *this deal however the* (3), *be allocated to deal* (3), *created a new deal* (3), *create a new deal* (3), *for this deal however* (3)
5	3	*record of a spot deal* (4), *should be allocated to deal* (3), *for this deal however the* (3)
6	1	*a record of spot deal* (4)
Total	**32**	

Both authors have *deal* n-grams that carry some technical meaning, such as *supply deal, loan deal, please match deal, the volume on this deal* and *for the term of the deal* for Germany, and *rolled deal, extended deal, corrected the deal* (discussed above), *a spot deal* and *allocated to deal* for Farmer. At the same time, however, a number of Germany's distinctive n-grams are not very specialised at all, such as *does this deal, this deal will be, the deal in the system* and *take a look at deal*. What is common to both Germany and Farmer is that there are more distinctive trigrams (three-word clusters) than any other length: 48 for Germany and 14 for Farmer. This suggests that n-grams of this size are most effective in capturing distinctive collocational patterns. It may be the case that the longer a lexical sequence is, the more likely it is to be idiolectal of the writer (Coulthard 2004; Culwin and Child 2010), but authors may be less likely to repeat these longer strings in their writing, rendering them difficult to use in attribution contexts. This appears to be the case here, as Germany and Farmer only have three and one distinctive *deal* six-grams respectively. Coulthard (2004: 440) argues that shorter lexical sequences are more likely (than longer ones) to be pre-assembled chunks of language made up of frequent collocations, as proposed by Sinclair's (1991) idiom principle. Therefore, drawing on this argument, and on the basis of the results for Germany and Farmer, it might be that trigrams are a key unit of psycholinguistic encoding. This argument is supported by Tomblin (2013: 102) and Larner (2014: 12), who found that when employing formulaic sequences as markers of authorship, there were more three-word clusters than any other length. Similarly, in his attribution of disputed texts in an asylum-seeker case, Juola (2013: 294) presents the results achieved using word trigrams as opposed to any other sequence length, or indeed any other linguistic feature (his JGAAP program is likely to have tested a wide range of features). As such, it may the shorter more common word trigrams that are more useful in attributing authorship. This is a hypothesis tested in Chapter 6.

The comparison of Germany's and Farmer's collocational profiles against the Base Rate Knowledge for the corpus indicated that these two authors used *deal* in distinctive ways. A closer qualitative analysis of a small number of collocates revealed collocational patterns that are distinctive of these two authors at population level, both within and beyond the Enron corpus. This was the first indication that this content word could provide idiolectal evidence. This subsequent analysis has identified a remarkable number of author-distinctive *deal* n-grams in the datasets for

Germany and Farmer. This further supports the argument for content words, or at least collocational frameworks in which content words appear. What this approach has done is pinpoint a pool of word n-grams that characterise and distinguish these authors' idiolects from the other authors in the Enron population. The prevalence of such distinctive clusters throughout their data is strong evidence to suggest that content words, and how authors use them, are fertile grounds for capturing idiosyncratic elements of language use. This analysis has focused on only *one* content word. Judging by the basis of how many author-distinctive n-grams have been identified, if the same analysis was run, identifying all distinctive two to six word clusters for every content word that each of these authors use, the number of author-distinctive word n-grams revealed would be enormous. By avoiding these words altogether on the basis that on their own they are over-dependent on topic, most stylometric analyses are not taking account of this vast range of author variation. This is regrettable, primarily because by analysing collocations and clusters in this way, we reveal more about authors' individual linguistic preferences and patterns than we do simply by counting relative frequencies of linguistic items. In forensic casework, the texts under analysis may be too short to apply any meaningful quantitative or statistical procedures. In such cases, the analyst would necessarily need to take advantage of all of the data available to them, and analysing content words and their collocations provides a means by which to do this.

In stylometric approaches to authorship analysis, content words have at best been avoided and at worst received unsubstantiated criticism over the last fifty years. Within a research tradition which focuses on relative frequencies of individual words this is not surprising, as on their own all they reveal about an author is that they write more or less about a particular topic than another author. However, moving beyond frequency and towards collocation analysis, the focus shifts into identifying the distinctive ways in which these content words are *used.* The argument that the analysis is too topic-sensitive is now less justified, because, as it has been shown here, different authors use the same content words and discuss the same topics in *different* ways. *Deal* is an intriguing example as it is central to the occupational roles of Germany and Farmer, and to the Enron community as a whole. Hoey (2005: 184–5) points out that language users will acquire stronger lexical primings associated with the topics, fields and contexts which they most frequently talk or write about. For Germany and Farmer, (and other Enron employees) their

occupations, their day-to-day language use, the topics that they discuss and the interactions that they have will all contribute substantially to their linguistic experience of, and lexical primings for, *deal*. What is important is that, as has been discovered, the lexical primings they have, at least insofar as the collocational patterns they produce, are idiolectal when tested against the population of Enron employees, all of whom are writing in the same community of practice and making very frequent use of the content word in question.

The analysis of *I* above revealed how a collocation analysis can identify potentially idiolectal evidence regarding the use of very common function words. This analysis of *deal* has made a case for focusing on content words in discussions of idiolect, by demonstrating how authors make idiolectal co-selections with them. The next section, and final one of this chapter, focuses on idiolectal variation within a specific type of speech act.

5.4 Idiolect in *please*-mitigated directives

There is a well-established relationship between the production of particular collocations or lexical sequences, and their functions in recurring and routine communicative situations. Sinclair (1991: 110), for example, suggests that the idiom principle may 'reflect the recurrence of similar situations in human affairs; it may illustrate a natural tendency to economy of effort'. Similarly, Nattinger and De Carrico (1992: 1) argue that 'just as we are creatures of habit in other aspects of our behaviour, so apparently are we in the ways we come to language'. They claim that routine formulae and prefabricated language chunks are products of 'ritualization' and play a large part in how we acquire and perform language. They also emphasize the importance of the relationship between lexical phrases and their function(s): 'their use is governed by principles of pragmatic competence, which also select and assign particular functions to lexical phrase units' (Nattinger and DeCarrico 1992: 36). Others have specifically investigated the role of lexical phrases in relation to their context. Becker (1975: 61) identifies certain lexical phrases as 'situational utterances', 'which are known to be the appropriate thing to say in certain circumstances'. Such phrases have been termed elsewhere as 'conversational routines' (Coulmas 1979; 81; Aijmer 1996) which are 'highly conventionalized prepatterned expressions whose occurrence is tied more or less to standardized communication situations' (Coulmas 1981: 2–3). Kecskés (2000: 606–7) uses

almost exactly the same definition for his 'situation-bound utterances', which he characterises as 'highly conventionalized, prefabricated pragmatic units whose occurrence is tied to standardized communicative situations', and whose 'use is highly predetermined by the situation'. More specifically, Wray (2002: 89) proposes that one of the main functions of formulaic sequences is 'to get the hearer to do something'. She continues: 'requests, demands, warning, orders and so on, are intended to get someone else to do something that we want for ourselves but cannot do personally'. She adds elsewhere that speakers use 'a range of markers (such as politeness markers) to frame them in a way that will maximise the likelihood of the required event coming about' (Wray 2000: 13). Schmitt and Carter (2004: 9) claim that 'because members of a speech community know these [formulaic] expressions, they serve a quick and reliable way to achieve the related speech act'.

Therefore, if such phrases, sequences and expressions are idiolectal for individuals, then when faced with the same kind of communicative situations, and when expressing the same type of speech act, individuals will produce different and distinctive linguistic output. Over the course of their lives, different authors will have developed and stored different phrases and collocational patterns for expressing the same speech acts. This is the hypothesis that is tested in this section, with particular focus on *please*-mitigated directives. *Please* is an important word in the Enron Corpus. It occurs a total of 11,061 times and is used by 165 of the 176 employees. It is the 24th most common word overall in the corpus, the second most common content word (after *thanks*), and the second most key word (also after *thanks*). A glance at its most common collocates in the corpus gives an indication as to what it is used for (Table 14).

Table 14. Top twenty *please* collocates in the Enron Corpus and EEC12

Enron Email Corpus			EEC12		
Collocate	**Authors**	**Freq.**	**Collocate**	**Authors**	**Freq.**
let	115	1,548	*review*	9	280
call	114	691	*let*	10	177
review	63	516	*call*	12	107
send	85	503	*send*	10	90
print	34	440	*see*	10	89
see	66	425	*take*	10	68
give	80	362	*forward*	10	66
forward	66	267	*print*	5	64
take	61	249	*put*	8	57
advise	42	229	*give*	10	53
add	63	227	*set*	9	52
make	73	217	*make*	10	47
put	47	203	*add*	10	32
contact	66	174	*advise*	8	32
get	65	168	*get*	9	32
note	52	168	*check*	7	29
set	45	154	*contact*	11	29
check	51	146	*note*	7	26
do	58	134	*keep*	6	21
find	43	131	*register*	1	21

All of these collocates are lexical verbs (except *do*, which can also be auxiliary), showing that employees are using *please* in either polite requests or mitigated directives, for example:

Example 41

<From: jim.steffes@enron.com>
<To: j.noske@enron.com>
<Subject: FW: Invitation: Please RSVP>

Please let the right person know that I will be on vacation.
Jim

Example 42

<From: gerald.nemec@enron.com>
<To: robert.walker@enron.com, eric.gillaspie@enron.com>
<Subject: Contract Goals>

Gentlemen, Yes that means you, Eric. **Please review** the attached and give me any comments before I distribute to the group.

Example 43

<Date: Tue, 18 Jan 2000 00:01:00 -0800 (PST)>
<From: richard.sanders@enron.com>
<To: twanda.sweet@enron.com>
<Subject: lunch>

Please call him and set this up

It is difficult to distinguish between requests and directives, given that they both serve the function of getting someone to do something. This study follows the distinctions between requests and directives as proposed by Searle (1969) and Bax (1968). The main difference between directives and requests is that 'the person who orders, as opposed to the requesting person, is necessarily in a position of authority over the recipient' (Searle 1969: 66). Similarly, Bax (1968: 676) states that in a request 'the requesting person benefits from the future act' and there is a 'reciprocal social relation' between the interactants. In a directive 'the person does not necessarily have to benefit from [the act]' and the addressee is in an 'inferior social relation'. Therefore, the use of *please* in Examples 41–43, and the uses of *please* focused on in this analysis are all directives. Directives with this kind of pragmatic purpose are strong contenders for producing recurring, pre-fabricated and formulaic collocations and phrases (Wray 2002: 89). As such, *please*-mitigated directives offer a speech act in which authors' potentially idiolectal linguistic output can be compared. Specifically, focus here will be on a selection of the most common *please*-mitigated directives in Enron, and how the twelve EEC12 authors make different lexico-grammatical choices when expressing the same speech act with the same pragmatic function. In the first instance, lexico-grammatical patterns are identified which distinguish between the twelve authors. These patterns are then tested against the whole of the Enron corpus to measure their population-level distinctiveness.

5.4.1 Different ways of saying the same thing

The types of variation that can distinguish between authors when they are expressing the same speech act can be grammatical or lexical. In the case of grammatical variation, the authors in question may use the same main verb, but do so with different complementation patterns. In the case of lexical variation, there may be a specific grammatical structure which is found across many authors, but the lexical choices which authors make within these structures serve to distinguish their styles.

An example of grammatical variation is with *please forward*. *Forward* is the eighth most common collocate of *please* in the corpus, occurring 276 times across the emails of 66 authors, and seventh most common in the EEC12, being used 66 times across ten of the twelve authors. It is used by authors of emails to request or command that their recipients send an email or attachment, either to them or a third party:

Example 44

<From: mark.haedicke@enron.com>
<To: john.novak@enron.com>
<Subject: Electrobolt Legal Opinion>

John:
I have not seen the legal opinion yet from Brazilian counsel [Dr. Pinto] on our dollar indexing issue. **Please forward** to me so I can review the opinion and get final approval for the deal.

Example 45

<From: vince.kaminski@enron.com>
<To: michael.sergeev@enron.com>
<Subject: Re: WTI CRUDE PRICE and NY Harbor Resid prices...>

Michael.
Thanks a lot for a very quick response. It looks fine.
Please, forward it to Margaret Carson. Please, explain the data source. She may expect NYMEX as the data source for natgas and WTI.

Example 46

<From: jim.steffes@enron.com>
<To: l.nicolay@enron.com>
<Subject: FW: PennFuture's E-Cubed - The $45 Million Rip Off>

Christi --
Some interesting language for the ICAP team. **Please forward** along.
Jim

This verb occurs in EEC12 with a relatively restricted range of complementation patterns; 60 of the 66 instances of *please forward* are followed by a noun phrase (NP) (29) or a prepositional phrase (PP) (31). The only recurring NP elements in the emails of the EEC12 authors are those with the determiners *the* and *your*, and those that are comprised of individual pronouns *it* and *this*. Even within this very small variation in complementation patterns, stylistic differences emerge between writers. Authors such as Kaminski and Farmer follow *forward* with NPs, creating the grammatical pattern *please forward* + NP + PP, including both the direct and indirect object:

please forward	NP	PP	
please, forward	***the*** *following message*	*to the Research Group*	(Kaminski)
please, forward	***the*** *message*	*to Dr. Kloucek.*	(Kaminski)
please, forward	***the*** *evaluation forms*	*to our guests?*	(Kaminski)
please forward	***this***	*to the appropriate person. Bill Bailey with*	(Farmer)
please forward	***this***	*to the appropriate person. Scott Berkman...*	(Farmer)
please forward	***this***	*to the appropriate person. We are buying*	(Farmer)
please forward	***this***	*to the appropriate person in your group.*	(Farmer)

In contrast, authors such as Nemec and Germany frequency omit the NP as direct object 22 and three times respectively, and so the grammatical pattern *please forward* + PP is more characteristic of their authorial styles:

please forward	PP	
please forward	***to*** *the appropriate accountant since Roger*	(Nemec)
please forward	***to*** *the appropriate individual.*	(Nemec)
please forward	***to*** *the appropriate individuals for review*	(Nemec)
please forward	***to*** *the appropriate persons at Greeley.*	(Nemec)
please forward	***to*** *all appropriate ROW agents.*	(Nemec)
please forward	***to*** *all appropriate land personnel*	(Nemec)
please forward	***to*** *whomever is appropriate*	(Nemec)
please forward	***to*** *Black Hills.*	(Nemec)
please forward	***to*** *AEC to facilitate their review.*	(Nemec)
please forward	***to*** *Oneok for their review*	(Nemec)
please forward	***to*** *the traders.*	(Germany)
please forward	***to*** *Mr. Storey's book admin.*	(Germany)
please forward	***to*** *Geoff Storey's book admin.*	(Germany)

These grammatical preferences for NP + PP or just PP serve to distinguish between these authors and their use of *please forward*. A number of these patterns retain distinctiveness at population level. *Please forward the*, with the definite article, as preferred by Kaminski is used 17 times by twelve of the other 175 different authors in the rest of the Enron corpus. Of these twelve authors, only three use it more than once, making this trigram relatively rare. Meanwhile, Farmer prefers *please forward this*, with the proximal deictic *this*, which as a trigram is relatively common in the rest of Enron corpus, occurring 33 times and being used more than once by eight additional authors. Every time Farmer uses this trigram, however, as shown in the annotated example lines above, it is part of the longer *please forward this to the appropriate person*, which is not used by anyone else in the Enron corpus. This particular phrase offers a contrast to Nemec, who also directs his recipients to

forward things to the appropriate people, but does so eliding *this* (as in the annotated examples above). Nemec's *please forward to the appropriate* is only found twice elsewhere in the corpus, both by one author. *Please forward to all appropriate*, however, which Nemec also repeats, is not used by anyone else in the corpus. Here we have two authors, Farmer and Nemec, who are giving exactly the same directive, but using different lexico-grammatical patterns to do so. Furthermore, these patterns are distinctive of these authors at population-level, providing potentially strong idiolectal evidence. Another point that should be noted is that in the examples above, Kaminski places a comma after *please* in clause-initial position. He does this regularly; he uses *please* 458 times, and in 402 (87.77%) of these it is followed by a comma. Clause-initial *please* followed by a comma is very rare in the rest of the Enron corpus, with only eight occurrences and one other person using it more than once.

Similar grammatical variation within the same phrase type is also found in the complementation patterns of *please see*. *See* is the sixth most common *please* collocate in the Enron Corpus, being found 425 times across the emails of 66 authors. It is the fifth most frequent collocate of *please* in the EEC12 sample, being used by ten authors and occurring 88 times. In 80 of its 88 (86.36%) instances it is being used by writers to direct recipients' attention to attachments or other messages:

Example 47

<From: james.derrick@enron.com>
<To: j.harris@enron.com>
<Subject: FW: National Bar Association sponsorship>

Please see the message below. Please let me have the materials when they arrive. Thank you.

Example 48

<From: phillip.allen@enron.com>
<To: jeffrey.hodge@enron.com>
<Subject: San Juan Index>

Liane,
As we discussed yesterday, I am concerned there has been an attempt to manipulate the El Paso San Juan monthly index. [...] **Please see** the attached spreadsheet for a trade by trade list and a summary. [...]

Example 49

<To: michele.sorensen@enron.com, richard.campbell@enron.com>
<Subject: California Strategy>
<Cc: jeremy.blachman@enron.com>

Michele & Rick --
Please see the note below. It is critical for you to call your clients (Jack in the Box/Burger King and Wendy's) to try and convince them to push their industry lobbyist to ask for a change in the date.

By far the most frequent complementation pattern of *please see* is *please see* + NP occurring 76 times. Despite being shared by ten of the twelve EEC12 authors, there is variation within this NP element which serves to distinguish individual authors' styles, as shown here:

please see	NP	
please see	***the attachment***. (x2)	(Derrick)
please see	***the attachment***.	(Farmer)
please see	***the attachment*** *below related to the OPM Hours Survey*.	(Farmer)
please see	***the attached*** *Tufco*	(Farmer)
please see	***the attached*** *Tufco file for July*	(Farmer)
please see	***the attached*** *spreadsheet related to the EEX supply*	(Farmer)
please see	***the attached*** *schedule.*	(Farmer)
please see	***the attached*** *memo.*	(Nemec)
please see	***the attached*** *guaranty*.	(Nemec)
please see	***the attached***. (x11)	(Nemec)
please see	***attached***. (x24)	(Nemec)

Derrick and Farmer both construct the noun phrase with the definite article and *attachment* as the head noun. Of the two, only Derrick does this more than once, and he is actually the only author in the whole Enron corpus to do so. Instead, Farmer prefers to use *attached* as an attributive adjective, premodifying *Tufco* (x2), *spreadsheet* and *schedule*. This is a feature he shares with Nemec, who uses *attached* to premodify *memo* and *guaranty*. This is very common in the Enron corpus; besides these two authors it is found 81 times in the Enron corpus, and eight authors use it more than once, some very frequently (Debra Perlingiere uses it 38 times). Much more often, Nemec chooses to use *attached* as the head noun, in *please see the attached*, omitting any further referent, which he does eleven times and is the only EEC12 author to do so. Aside from Nemec's uses, this *please see the*

attached followed by a full stop is only found an additional eleven times in the corpus, being used more than once by only three authors. Furthermore, on 24 occasions Nemec also omits the definite article, writing the telegraphic *please see attached*. Again, he is the only EEC12 author to use this three word cluster, and it is rare in the rest of the Enron corpus, occurring only 25 times and being used by only four authors more than once. In the same way as with *please forward*, then, although many authors can share the same overall complementation pattern, there may still be grammatical variation within the complement element that serves to distinguish particular authors' styles from those of others, even at population level.

Variation can be found in the ways in which the EEC12 authors tell their assistants to put an event or meeting on their calendars. The most important verb in this speech act is *put*. *Put* is the thirteenth most frequent *please* collocate in the Enron corpus, occurring 203 times across the emails of 47 authors. Of these 203 occurrences, *please put* refers to *calendar* or *schedule* 107 times (52.71%). In the EEC12, 45 of the 57 (78.95%) instances of *please put* refer to *calendar* (sometimes spelt *calender*) or *schedule*:

Example 50

<From: mark.haedicke@enron.com>
<To: janette.elbertson@enron.com>
<Subject: Friday's US Regulatory Conference Call>

Please put this on my calender.

Example 51

<From: vince.kaminski@enron.com>
<To: shirley.crenshaw@enron.com>
<Subject: FW: Market Maker Discussion>

Shirley,
Please, put it on my schedule.
Vince

Please put appears with two complementation patterns in the EEC12; *please put* + NP + PP and *please put* + PP. The only authors who used *please put* + NP more than once are Kaminski (n=15) and Haedicke (n=5), and they both consistently fill the NP slot with a pronoun. However, what distinguishes their styles is that Kaminski always uses *it*, while Haedicke prefers *this*:

please put	NP	PP	
please, put	***it***	*on my schedule.* (x9)	(Kaminski)
please, put	***it***	*on my calendar.* (x6)	(Kaminski)
please put	***this***	*on my calendar/calender* (x3)	(Haedicke)
please put	***this***	*on my schedule*	(Haedicke)
please put	***this***	*tentatively on my schedule*	(Haedicke)

Haedicke uses proximal deictic *this* to refer to the meeting or event in question, as referred to in the subject field of the email. In Example 50 above, Haedicke's *this* refers to 'Friday's US Regulatory Conference Call' as stated in the subject line. *Please put this on my* is used a total of 19 times in the Enron corpus besides Haedicke's four instances, and is found more than once in the emails of only three authors. As is also shown in Example 50, Haedicke misspells *calendar* as *calender*, which he does twice out of three uses. In turn, the six-gram *please put this on my calender* (misspelt) is only found once elsewhere in the corpus. In comparison with Haedicke's proximal deictic *this*, Kaminski consistently prefers the pronoun *it*, as in Example 51. The resulting phrase *please put it on my schedule* appears only once outside of Kaminski's emails, while *please put it on my calendar* followed by a full stop does not occur anywhere else in the corpus. In addition to these forms Haedicke and Kaminski, along with Steffes, both omit the NP and object:

please put	PP	
please put	*on my <u>calendar</u>*	(Haedicke)
please put	*on my <u>schedule</u>* (x2)	(Haedicke)
please, put	*on my <u>calendar</u>* (x2)	(Kaminski)
please, put	*on my <u>schedule</u>* (x9)	(Kaminski)
please put	*on my <u>calendar</u>* (x11)	(Steffes)

Haedicke and Kaminski are the only two authors in the Enron corpus who use *please put on my schedule*. Kaminski (n=9) uses it more frequently than Haedicke (n=2) and their usage can be distinguished by Kaminski's consistent use of a comma after *please*. As such, *please, put on my schedule* (with comma) and *please put on my schedule* (no comma) are unique to Kaminski and Haedicke respectively in the Enron corpus. Kaminski also uses *please put on my calendar* (as does Haedicke, but

only once), and this is a feature he shares with Steffes. Steffes' use of this phrase can be distinguished from that of both Kaminski and Haedicke, however, in that whereas they alternate between the *please put* NP + PP and *please put* + PP patterns, he always omits the NP. In addition, he is also consistent in his use of the noun *calendar* rather than *schedule*, which Haedicke and Kaminski also alternate between. Again, Kaminski's use of this phrase is different from Haedicke's and unique in the whole corpus as both times he follows *please* with a comma. *Please put on my calendar* with no comma, which Steffes uses eleven times, is only found 13 times in the rest of the corpus and is used more than once by four authors. Haedicke and Kaminski's use of *schedule* to express this speech act is not restricted to its noun form. They both, along with Arnold, use *schedule* as a verb. What distinguishes Haedicke's style in this case is that he is the only one of the three authors to use *please schedule* intransitively, followed by a full stop and with no explicit reference to the event or meeting in question, which he does five times:

please schedule	NP	
please schedule	*a round of interviews with john griffith...*	(Arnold)
please schedule	*me (for 10/10 from 3:00-5:00.)*	(Arnold)
please schedule	*an interview with this guy.*	(Kaminski)
please schedule	*a meeting, 30 minutes with him.*	(Kaminski)
please schedule	*a meeting on Thursday, Aug 17.*	(Kaminski)
please schedule	*an interview with Konstantin on May 8*	(Kaminski)
please schedule (x5)	[intransitive]	(Haedicke)
please schedule	*30 minutes with Janette.*	(Haedicke)
please schedule	*this call.*	(Haedicke)
please schedule	*something either in my office or a lunch.*	(Haedicke)
please schedule	*a meeting for Dan Fournier and David...*	(Haedicke)
please schedule	*lunch in late April*	(Haedicke)

This intransitive *please schedule* form is used only by Haedicke in the EEC12 sample, and occurs four times in the wider corpus, three of which are by the same author (Steven Kean).

Besides shared complementation patterns, there are cases in the Enron Email Corpus in which authors produce the exact same grammatical structure when expressing particular speech acts, but it is their choice of synonymous or alternative verbs within these grammatical patterns which distinguishes them. An example of this is different authors' preferences for *please call, please contact* and *please give*

me a call. *Call, contact* and *give* are all in the top twenty *please* collocates in both the Enron Corpus as a whole and the EEC12 sample. Of the 362 instances of *please give* in the Enron Corpus, *a call* is the direct object of the verb 231 (63.81%) times and this is the case for 26 of the 53 (49.06%) occurrences of *please give* in EEC12. The subsequent *please give me a call*, is a delexicalised verb (Stubbs 2001: 32). Therefore, these three phrases are different ways of expressing the same speech act, which authors use to request or demand that the recipient call them or someone else:

Example 52

<From: chris.germany@enron.com>
<To: hattie.golden@williams.com>
<Subject: Crystal Ballroom>

I would like to RSVP for the Houston Energy Expo Customer Celebration in The Crystal Ballroom in the Rice Hotel on Wed, March 21, 2001. **Please call** me at ***-***-**** if you have any questions.

Example 53

<From: gerald.nemec@enron.com>
<To: randall.curry@enron.com>
<Subject: Revised Unocal Energy Trading Guaranty>

Attached for your execution and distribution is the amended and restated Unocal Energy Trading, Inc. guaranty. The amount has been increased from $5 mil to $10 mil and Union Oil Company of California has been added as a counterparty. If you have any questions, **please contact** me at ***-****.

Example 54

<From: darren.farmer@enron.com>
<To: jeff.hodge@enron.com>
<Subject: Cleburne Plant - Tenaska IV>
Jeff,
I left a voice mail message for you yesterday about our role as agents for the Tenaska IV's Cleburne plant.[...] **Please give me a call** when you get a chance (3-6905).

There are differences in the frequencies with which the twelve EEC12 authors use these three verbs (Table 15).

Table 15. *Please call, contact* and *give * a call* in EEC12

Author	**Total**	***please call* (n=107)**		***please contact* (n=29)**		***please give * a call* (n=28)**	
Arnold	**5**	4	(80.00)	1	(20.00)		
Germany	**22**	20	(90.91)	2	(9.09)		
Lavorato	**14**	8	(57.14)			6	(42.86)
Zipper	**3**	1	(33.33)			2	(66.67)
Allen	**10**	9	(90.00)	1	(10.00)		
Dorland	**2**	1	(50.00)	1	(50.00)		
Farmer	**7**	2	(28.57)	2	(28.57)	3	(42.86)
Kaminski	**41**	24	(58.54)	9	(21.95)	8	(19.51)
Derrick	**5**	3	(60.00)	2	(40.00)		
Haedicke	**10**	4	(40.00)	2	(20.00)	4	(40.00)
Nemec	**24**	13	(54.17)	9	(37.50)	2	(8.33)
Steffes	**21**	18	(85.71)			3	(14.29)

However, these differences in preference are illuminated more clearly by authors' choices within recurring grammatical structures. An example is the grammatical pattern *please* + V + NP + PP, often containing *at* and their telephone number (asterisked here):

please	V	NP	PP	
please	***call***	*me*	*at* ***-***-****...	(Germany)
please	***call***	*me*	*at* ***-***-****...	(Germany)
please	***call***	*him*	*at your convenience*	(Germany)
please	***call***	*Beverly*	*at* ***-***-****...	(Germany)
please	***call***	*her*	*at home.*	(Germany)
please,	***call***	*me*	*at the end of the day*	(Kaminski)
please,	***call***	*him*	*at the number* **********...	(Kaminski)
please	***call***	*me*	*at x-*****.	(Nemec)
please	***call***	*me*	*at* ***-****.	(Nemec)
please	***call***	*me*	*at (***)* ***-****.	(Nemec)
please	***call***	*me*	*at (713)* ***-****.	(Nemec)
please	***call***	*me*	*at x-***** with any questions.*	(Nemec)
please	***call***	*me*	*at your convenience.*	(Nemec)
please	***call***	*me*	*at* ***-***-****.	(Steffes)
please	***call***	*me*	*at* ***-***-**** to assist...*	(Steffes)
please	***call***	*me*	*at 3x****	(Steffes)
please	***contact***	*me*	*at (***)-***-****.*	(Nemec)
please	***contact***	*me*	*at* ***-***-****.*	(Nemec)
please	***contact***	*me*	*at (***)-***-****if you would*	(Nemec)
please	***contact***	*me*	*at x**** if you would*	(Nemec)
please	***contact***	*me*	*at (***)* ***-****.*	(Nemec)
please,	***give me a call***		*at* ***-***-**** and we can...*	(Kaminski)
please,	***give me a call***		*at your convenience and we...*	(Kaminski)

The verb (V) slot in this pattern is most commonly filled by *call*, with Germany, Kaminski, Nemec and Steffes all using this more than once. In comparison, in addition to using *call*, Nemec is the only author of the twelve in EEC12 to use *contact* in the verb position more than once, producing *please contact me at* five times. This phrase is rare in the Enron corpus as a whole; besides Nemec's five occurrences, it only appears four times, once in the emails of four different authors. In contrast, Kaminski, though only using it twice, is the only EEC12 author to use *please give me a call at* more than once. This is more common in the Enron population, being found an additional 18 times, yet still only two authors use it more than once. Here, the grammatical pattern of *please* + V + NP + PP is shared across all of the authors, but it is their choice of verb that distinguishes them from one another. This kind of lexical variation is also found within the grammatical structure *please* + V + NP + *when* adverbial clause. Steffes and Germany use *please call me when*, as opposed to Farmer, who prefers *please give me a call when:*

please	V	NP	*when* adverbial	
please	***call***	*me*	*when you get some time*	(Steffes)
please	***call***	*me*	*when you get in.*	(Steffes)
please	***call***	*me*	*when you get a chance.*	(Germany)
please	***call***	*me*	*when you get a chance at* ***-***-****.	(Germany)
please	***call***	*me*	*when you get a chance to discuss everything*	(Germany)
please	***give me a call***		*when you get a chance.*	(Farmer)
please	***give me a call***		*when you get a chance (*-*****).*	(Farmer)

A clear distinction can be drawn, here, between Germany and Farmer, who both tell their recipients to contact them 'when you get a chance', but choose different verbs in doing so. Germany uses *please call me when you get a chance* three times, and it is found in the data of only one other author in the Enron corpus, Kim Ward, who uses it twice. In contrast, Farmer uses *please give me a call when you get a chance,* which appears only once in the rest of the corpus. Finally, the same is the case with the longer structure of *please feel free to* + V + NP. This pattern is used by three of the twelve EEC12 authors: Kaminski, Steffes and Nemec. While all three authors write *please feel free to call*, only Kaminski in the EEC12 uses *please feel free to contact.* At population level, this phrase is used an additional 17 times, and three authors use it more than once.

please	*feel free to*	V	NP		
please	*feel free to*	***call***	*him*	*with any question*	(Kaminski)
please	*feel free to*	***call***	*Mike*	*and discuss this matter with...*	(Kaminski)
please	*feel free to*	***call***	*me*	*if this isn't getting fixed*	(Steffes)
please	*feel free to*	***call***	*me*	*at 3x7673.*	(Steffes)
please	*feel free to*	***call***	*me.*		(Nemec)
please,	*feel free to*	***contact***	*her*	*and give my name as a…*	(Kaminski)
please,	*feel free to*	***contact***	*her*	*directly and discuss the…*	(Kaminski)
please,	*feel free to*	***contact***	*Ed*	*directly and let me know…*	(Kaminski)
please,	*feel free to*	***contact***	*me*	*at 713 853 3848.*	(Kaminski)

This detailed analysis of four major *please*-mitigated directive types across the twelve authors in the EEC12 sample has identified a number of lexico-grammatical patterns which are distinctive of individual authors, not only within this sample, but at population level when compared with the rest of the Enron corpus. This provides evidence to support the hypothesis that people express the same speech act in different ways. In turn, the different realisations of the same speech act across authors represent clear manifestations of the distinct idiolects and linguistic preferences of these authors. The following section summarises the findings of this analysis, and discusses their implications for a theory of idiolect.

5.4.2 Author-distinctive *please* n-grams

The focus of this analysis has been on four major types of directives in the Enron corpus: instructing recipients to forward messages, directing recipients' attention to attachments, ordering them to put a date on a calendar, and telling recipients to contact them or someone else. These were identified as frequent directives on the basis that the verbs *forward, see, put* and *call/contact/give* were very frequent collocates of *please* in the Enron Email Corpus. The frequent recurrence of these four directives indicates that they are central to the day-to-day communication within Enron emails. The routine nature of these directives, and the familiar communicative contexts in which they arise, make them fertile environments for the production of lexical phrases (Nattinger and DeCarrico 1992), formulaic sequences (Wray 2002), situation-bound utterances (Becker 1975–61; Kecskés 2000: 606–7) or conversational routines (Coulmas 1979; 81; Aijmer 1996). The theory behind these very similar constructs, as applied in this analysis, is that because Enron employees repeatedly find themselves in the same communicative situations on a daily basis,

they develop phrases and lexical sequences which reduce the cognitive load of production and which they know through experience are effective in expressing the speech act required. In turn, this analysis set out to investigate the extent to which the resultant lexical sequences are idiolectal.

Across the four directives examined, there are 23 word n-grams which have been identified as being distinctive of individual authors (Table 16).

Table 16. Author-distinctive *please* n-grams in the EEC12 sample

n-gram	**author (freq)**	**Enron Freq.**	**Enron authors***
please,	Kaminski (402)	8	1
please forward the	Kaminski (3)	17	3
please forward this	Farmer (4)	33	8
please forward this to the appropriate person	**Farmer (4)**	**0**	**0**
please forward to the appropriate	Nemec (4)	2	1
please forward to all appropriate	Nemec (2)	0	0
please see the attachment	**Derrick (2)**	**1**	**0**
please see the attached.	Nemec (11)	11	3
please see attached.	Nemec (24)	25	4
please put this on my	Haedicke (4)	19	3
please put this on my calender	**Haedicke (2)**	**1**	**0**
please put it on my schedule.	**Kaminski (9)**	**1**	**0**
please put it on my calendar.	**Kaminski (6)**	**0**	**0**
please put on my schedule	**Haedicke (2)**	**0**	**0**
please, put on my schedule	**Kaminski (9)**	**0**	**0**
please, put on my calendar	**Kaminski (2)**	**0**	**0**
please put on my calendar	Steffes (11)	13	4
please schedule.	Haedicke (5)	4	1
please contact me at	**Nemec (5)**	**4**	**0**
please give me a call at	Kaminski (2)	18	2
please call me when you get a chance	Germany (3)	2	1
please give me a call when you get a chance	**Farmer (2)**	**1**	**0**
please feel free to contact	Kaminski (4)	17	3

*Number of other Enron authors who use this n-gram more than once

Some of these have been found to mark different lexical choices between the EEC12 authors (such as *please contact me at* versus *please give me a call at*) and others which mark different grammatical preference (such as *please forward this to* versus

please forward to), including patterns of ellipses. Furthermore, when tested against the full Enron Email Corpus, most of these are found to be distinctive at population level. When Enron authors are faced with similar communicative situations in which they are required to express the same speech act, they produce different linguistic output. So, although they are all routinely part of similar situations in which they need to direct their recipient to *forward, see, put* and *call/contact/give,* they have developed different phrases for doing so. In turn, this offers empirical evidence to support the theory of idiolect. As with *deal* in Section 5.3.2 the word n-grams in Table 16 capture observable and quantifiable segments of authors' idiolects. More specifically, they capture the distinctive lexical *co-selections* authors make with *please*. A question arises here as to how distinctive an n-gram needs to be to be considered evidential of an idiolect. Those which provide perhaps the strongest idiolectal evidence are those which are either not found at all in the rest of the Enron corpus, or are not used more than once by any other author in the corpus (bold in Table 16). That said, none of the 23 word n-grams are used more than once by more than eight additional authors in the Enron corpus. Out of 176 authors from within a relevant population of writers from the same linguistic community, an n-gram that is shared by only eight authors is rare enough to be considered distinctive. Another point to consider is that, with the exception of Kaminski's *please* followed by a comma, most of these word n-grams are relatively infrequent; the most frequently occurring distinctive n-gram is Nemec's *please see attached.,* which he uses 24 times. On the one hand, this may be problematic for forensic authorship analysis. If disputed texts are short, and known documents for candidate authors are limited, then features which authors use infrequently may be more unlikely to appear, and therefore more unlikely to be useful in attributing authorship. On the other hand, the infrequency of distinctive word n-grams may give us an insight into the nature of a person's idiolect that we are able to identify. To expect a person's idiolect to be manifest in writing by extremely frequent and prominent patterns which stand out as being unique against any given Base Rate Knowledge is unrealistic. Instead, it may be rarer linguistic idiosyncrasies and less prominent patterns that are true markers of idiolect. Indeed, Mollin (2009: 387) identifies the maximiser collocation *absolutely frank* as being idiolectal of Tony Blair, despite him using it only 19 times in a three million word corpus. Similarly, as Koppel and Schler (2003: 6) note, lexical and syntactic features 'never quite disappear from view, but they are rarely used with

such outlandish frequency as to serve as smoking guns' of idiolect and authorship. Indeed, such elements of idiolect may be too infrequent to use in the stylometric approaches which they and others use. A corpus linguistic approach, however, combining qualitative and quantitative evidence has identified such lexico-grammatical variation as being indicative of idiolect. In forensic psychology and behavioural linkage across crimes, there is an understanding that 'rare characteristics may be important for linking offences to one individual' (Canter and Heritage 1990: 191). In the linking of behaviours across rape cases, for example, 'the more rare the behaviors reported in rape, the more likely they were to contribute to distinct varieties of sexual assault' (Canter et al. 2003: 159). In a linguistic context, it might be the more rare the linguistic behaviours, the more likely they are to belong to a distinct idiolect, and therefore the more important they might be in linking texts to one individual.

5.5 Chapter conclusion: corpus, collocation and idiolect

It is being increasingly acknowledged in forensic authorship attribution, from both stylometric and stylistic camps, that it is time to move beyond the situation in which the two methodologies are competing, and that focus should be on how they can complement one another. One of the most effective ways of combining these approaches is by: (i) identifying linguistic features for which their variation across authors can be explained theoretically, and (ii) using them in accurate and reliable statistical attribution procedures. This chapter has addressed the first of these points.

As the conceptualisation of 'idiolect' shifts away from the abstract notion it has traditionally been, forensic linguists are now beginning to focus on how individuals' idiolects are manifest in actual language production, and how distinctive this language production is at 'population-level' when compared against the Base Rate Knowledge of a relevant linguistic community. (Grant 2010; 2013; Turell and Gavaldà 2013). In this chapter, the Enron Email Corpus has been used as dataset against which the population-level distinctiveness of individual employees' linguistic patterns can be measured. In Turell and Gavaldà's (2013: 499) terms, it is 'a relevant population' of writers 'from the same linguistic community'. It represents the daily communicative practices of 176 employees of the same company, all of whom are communicating using the same medium, ultimately towards achieving the same institutional goals. The desire for a 'relevant

populations' of writers has gained a lot of momentum in recent research and commentary (e.g. Kredens 2002; McMenamin 2010; Kredens and Coulthard 2012; Butters 2012; Grant 2013) (see Section 2.1). Forensic linguists see the clear value in comparing disputed documents in forensic cases with documents written by similar types of people, using the same medium and genre around the same period of time. However, the use of a 'relevant population' is almost invariably and inevitably synonymous with using a 'smaller population' (than *all* speakers of the given language). This may not be problematic in identifying idiolect *per se*. However, it does mean that the quantitative conclusions that forensic linguists can possibly make about the rarity of features may be less reliable (or at least less impressive) than those of DNA and fingerprint experts, and even forensic phoneticians, who have at their disposal far larger reference data or more 'general' populations. Therefore, the findings here can be qualified as being valuable to the extent that the Enron Email Corpus can be considered as representing a 'linguistic community', and the quantitative results of a 'relevant population' can be considered reliable.

The corpus-based approach to identifying idiolectal language use employed here can be applied in forensic casework to reveal author-distinctive lexico-grammatical patterns. With machine-readable data and a suitable piece of computer software (*Wordsmith, AntConc, SketchEngine*), this methodology can be replicated with relative speed and ease. Word lists and keyword lists can be performed to instantly identify salient lexical choices in any data, and simple concordance searches will show the use of individual words or phrases in context. As with any linguistic analysis, the linguist will be required to identify and interpret emerging patterns, but the corpus software makes the analysis much less time consuming and labour intensive (and more reliable) than traditional non-corpus approaches. These are all attractive features when operating under the time-pressures of forensic work.

This chapter has found that Enron authors exhibit distinctive collocational patterns and word n-grams with very common and high frequent lexical items: *I, deal,* and *please*. The importance of these words in particular is that they demonstrate that even within lexical items and linguistic practices that are widely shared amongst a linguistic community, distinctive linguistic preferences can be identified. The results here, therefore, offer empirical evidence in support of a theory that collocational patterns are idiolectal for individual language users. The implication of this for authorship analysis is that collocations are an aspect of

language use that are underpinned by a theory of idiolect. Variation in collocation patterns between authors can be explained by the argument that each individual has a unique set of linguistic interactions, experiences and memories, on the basis of which they build unique associations and relationships between words. Therefore, authors' productions of these collocations and lexical sequences in their writing are identifiable realisations of their idiolects. Furthermore, these idiolectal collocation patterns can be isolated and quantified as word n-grams (strings of *n* words). The next chapter, therefore, shifts focus to point (ii) above, and uses these word n-grams to capture idiolectal collocations and lexical sequences and incorporates them into a statistical authorship procedure.

6 Attributing authorship using word n-grams

The main drawback in stylometric authorship research is that there is rarely any theoretical motivation behind the linguistic features used to distinguish between authors, and so the results produced are difficult to interpret or explain in linguistic or stylistic terms (Grant 2008: 226; Howald 2008: 235; Argamon and Koppel 2013: 299; Stamatatos 2013: 428) (see Section 2.2.3). The aim of this chapter is to develop a methodology for authorship analysis that combines the accuracy and statistical reliability of stylometric techniques with the linguistically-motivated analyses of stylistic approaches. The previous chapter has provided empirical evidence to suggest that collocation patterns and lexical sequences are idiolectal for individual speakers and that these patterns can be isolated and quantified as word n-grams. Drawing on word n-grams as features for analysis, it is possible to develop an approach for authorship attribution which is based on the statistical measurement of linguistic similarity between texts, and which provides results that are underpinned by a theory of idiolect. Such a method, if successful, would represent a significant move towards bridging the theoretical and methodological gap between the two divergent approaches to authorship attribution.

The discussion of the results reported here answers three research questions:

1. How successful is the method of using word n-grams in correctly identifying the author of disputed email samples from a pool of 176 Enron employees?
2. How robust is the method in relation to changes in the size of the disputed samples being attributed?
3. Which length of n-gram is best for attributing authorship of Enron email samples?

After these questions have been addressed, a case study of one individual will examine how linguistic results can be used to explain and underpin the statistical results. To conclude, this approach is discussed and evaluated in terms of its contribution towards bridging the gap between stylometric and stylistic methodologies.

6.1 The approach

In this experiment, each of the twelve EEC12 authors have ten different random samples of five different sample sizes (2%, 5%, 10%, 15% and 20%) extracted from their email set. In each test in this experiment, one extracted sample at a time is compared against the remaining emails (either 98%, 95%, 90%, 85% or 80%) of the author to whom the sample belongs, and the whole email sets of the other 175 candidate authors. In forensic casework terms, the extracted samples represent the 'disputed' documents, the author of which is to be identified, and the texts against which these samples are compared represent the 'known' writings of (in this case 176) possible authors. The basis of comparison between the disputed samples and the known texts in this experiment are the word n-grams between one and six words, for example:

unigram	–	*please*	four-gram	–	*please format and print*
bigram	–	*please format*	five-gram	–	*please format and print the*
trigram	–	*please format and*	six-gram	–	*please format and print the attachment*

The use of these word n-grams differs from other studies that have used word sequences (Hoover 2002; 2003; Coyotl-Morales et al. 2006; Sanderson and Guenter 2006; Grieve 2007; Juola 2013) in two main ways. First, in this experiment, *all* of the word n-grams in the data are utilised, not just the most frequent ones or ones which begin with or include a particular node word. The advantage of this approach is that there is no subjectivity with regard to which n-grams are or are not included in the analysis. It also ensures that the comparisons utilise all of the data available in both the disputed and known data. Second, on the basis of studies such as Coulthard (2004: 441) and Culwin and Child (2010: 16), and the collocation analysis in the previous chapter, word n-grams of up to six words in length are tested, rather than only two- or three- word sequences that are more commonly used. The similarity between the disputed and known texts is based on the number of word n-grams that are shared between the two datasets being compared as a proportion of all the data in the sets combined. This similarity is measured using the Jaccard similarity co-efficient (Section 4.4.3). The Jaccard statistic takes into consideration only whether a particular n-gram is found in both the disputed sample and known texts, rather than how frequently it occurs. The entire process of sample extraction, text comparison and Jaccard calculation are performed using *CFL Jangle*, and the full procedure is detailed in Section 4.2 above.

Given that each of the twelve authors has ten random samples of each of the five sizes, that gives a total of 600 disputed samples being subject to testing in this experiment. Furthermore, each of these samples is tested for their similarity against the 176 known comparison files (one for the real author and one each of the other 175 authors) using the six different n-grams (1–6 words). This means that, overall, this experiment comprises 3,600 individual tests: a substantial, systematic, empirical testing of this method. Usefully for forensic applications, once the data in question has been cleaned up and prepared for analysis, an individual attribution test using *Jangle* takes a matter of seconds. In this chapter, the accuracy and reliability of the method is evaluated in the following ways:

1. Raw attribution accuracy. The first most straightforward way of assessing success is to consider the number of times individual samples are attributed to their correct author. That is, if the extracted 'disputed' sample of emails achieves the highest Jaccard score of similarity when tested against the set of remaining emails by the same author, then attribution has been successful. If a set of emails written by one of the other 175 authors achieves the highest Jaccard score, then attribution has been unsuccessful. This is done for all twelve authors, with all sample sizes, using all six n-gram types.
2. Mean Jaccard score. The second way involves considering the mean Jaccard scores obtained by all 176 candidate authors over the ten tests for each sample size using the different n-gram types. This is an important measure given that, although the actual author of individual samples may not always achieve the highest Jaccard score in an individual test, they may achieve Jaccard scores consistently high enough that they have the highest mean Jaccard score over the ten tests for that sample size of all 176 candidate authors. In such a case, attribution of the ten samples for that author would be considered successful.

6.2 The samples

The authors whose emails are sampled and used in the authorship attribution experiment are the twelve who comprise the EEC12: Allen, Arnold, Derrick, Dorland, Farmer, Germany, Haedicke, Kaminski, Lavorato, Nemec, Steffes and Zipper. For each of the twelve authors, ten samples of 20%, 15%, 10%, 5% and 2% of their emails were extracted, each sample random and each one different. 20% of emails was chosen as the maximum threshold of data to be attributed, after the pilot study found near perfect results using unigrams on samples of this size (explained in Section 4.5). The decision to have the attributable unit as complete emails rather than sets of a specific number of tokens followed previous authorship studies which have concerned themselves with identifying the authors of full texts or full messages

(e.g. Chaski 2001; Grieve 2007; Grant 2007; Grant 2013; Nini and Grant 2013). Had a specific number of tokens (or series of tokens) been used instead, this would have split some emails in two, with the 'questioned' datasets containing parts of emails, and the 'known' sets containing the remainder. Using full emails avoids this. Also, full emails are units which can remain constant across all authors. This means that because different authors have different amounts of data and write emails of different lengths, selecting the same proportion of emails (2%–20%) across authors produces a much wider range and variety of dataset sizes on which to test the effectiveness of method. That said, it is useful for forensic casework to be able to extrapolate the results in terms of tokens, and so the mean number of tokens that make up each sample size for each of the twelve authors are given in Table 17. The reason behind taking so many samples of each size was to assess consistency and reliability of results. This approach is similar to the machine learning technique known as '(ten-fold) cross-validation', in which equal-sized subsets of the data are used train the algorithm ('training sets'), and the rest of the data is used to test it ('test set'), with the accuracy of the method being evaluated through the average success rate across the various tests (Argamon and Koppel 2013; Koppel et al. 2013). Although not identical to this, the use of ten different randomised samples of each size in the experiment here increases the confidence we can have in the accuracy and reliability of the results obtained.

Given that the twelve authors of EEC12 have email sets of vastly different sizes, the samples extracted also vary in size across the authors (Table 17). For example, the 20% samples range from 48 emails (Zipper) to 459 (Germany and Kaminski), and a mean of 951 tokens (Derrick) to 13,436 (Germany). In fact, the actual range for the 20% samples is from Derrick's smallest sample of 762 tokens, to Germany's largest of 14,859, the latter of which is the largest sample in the entire experiment. At the smallest end of the spectrum, the 2% samples range from as few as four emails (Zipper) to 45 (Germany and Kaminski), and a mean of 89 tokens (Derrick) to 1,317 (Germany). The smallest sample in the experiment is Derrick's of only 55 tokens. This coverage of the five different sample sizes across the twelve authors ensures that this method is applied to a very wide range of 'disputed' dataset sizes. This is particularly important in potential forensic applications of this method, as the disputed documents of which the authorship is questioned are often unhelpfully short (Coulthard and Johnson 2007: 172; Cotterill 2010: 578).

Table 17. The sizes of disputed samples in the authorship attribution experiment (mean tokens across the ten samples for each size).

	20%		15%		10%		5%		2%	
	Emails	**Tokens**	**Emails**	**Tokens**	**Emails**	**Tokens**	**Emails**	**Tokens**	**Emails**	**Tokens**
Germany	459	13,436	344	10,116	229	6,820	114	3,367	45	1,317
Kaminski	459	9,362	344	6,909	229	4,656	114	2,200	45	955
Nemec	293	9,224	219	7,047	146	4,963	73	2,327	29	998
Steffes	240	6,642	180	5,126	120	3,033	60	1,400	24	655
Lavorato	222	4,600	166	3,360	111	2,217	55	1,031	22	450
Arnold	207	4,633	155	3,281	103	2,481	51	1,272	20	449
Haedicke	160	3,294	120	2,309	80	1,591	40	780	16	286
Farmer	154	4,208	115	3,213	77	2,016	38	1,030	15	350
Dorland	100	2,464	75	1,935	50	1,279	25	655	10	280
Derrick	93	951	70	712	46	485	23	253	9	89
Allen	71	2,575	53	2,001	35	1,267	17	609	7	255
Zipper	48	1,281	36	1,090	24	554	12	259	4	151

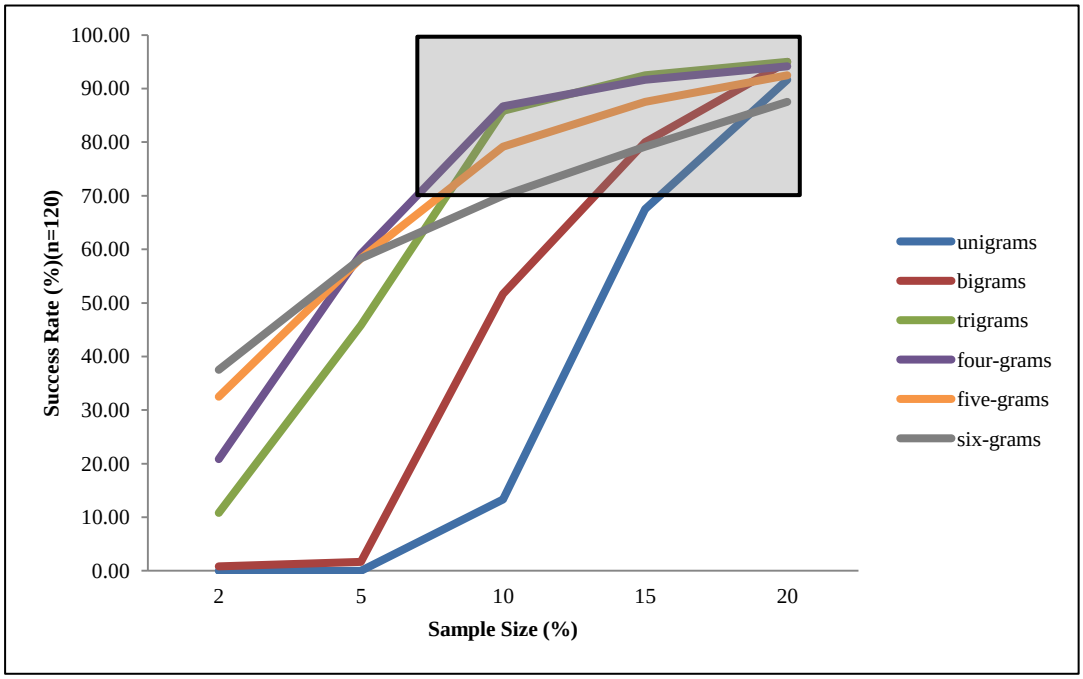
Of these samples, only the 20% sets of Germany, Kaminski and Nemec approach anything like the 10,000 words required for a reliable authorial set (Burrows 2007: 30). Therefore, even the largest sample sizes used in this study may be considered relatively small when compared with those of other computational studies. Such studies have commonly employed attributable samples of between 1,000 and 39,000 tokens (e.g. Hoover 2004; Argamon and Levitan 2005; Burrows 2005; Labbé 2007; Savoy 2012). At the same time, the 2% samples in this study of between a mean of 89 and 1,317 tokens are comparable to those used in studies which have tested with smaller samples of between 140 and 1,300 tokens (van Halteren et al. 2005; Koppel et al. 2011, 2013; Hirst and Feiguina 2007; Luyckx and Daelemans 2011), and those in a forensic context that have used exceptionally small test sets of between 105 and 543 tokens (Chaski 2001; 2005; Grant 2007; Rico-Sulayes 2011). Similarly, Grant (2007) reduces his question sets to as small as three texts, and the 2% samples of Dorland, Derrick, Allen and Farmer represent similar sized sets. Furthermore, as well as these sample sizes being relatively—if not exceptionally—small by computational or stylometric authorship standards, a pool of 176 candidate authors is large in relation to other experimental authorship attribution research. While there are exceptions, such as Koppel et al. (2011) and Narayanan et al. (2012) who use open candidate sets of thousands of potential authors, most quantitative authorship studies have tested methods on much smaller candidate sets of three (Grant 2007), six (Juola 2013), 10 (Rico-Sulayes 2011: 58–9), 20 (Zheng et al. 2006: 387), 40 (Grieve 2007: 258), and 145 (Luyckx and Daelemans 2011: 42). Overall, the combination of small sample sizes and a large number of candidate authors makes the attribution task in this study a relatively difficult one.

6.3 Results

6.3.1 Overall accuracy

Figure 25 presents the raw attribution accuracy of all six n-grams across all five sample sizes. Of the 3,600 tests in this experiment, 2,120 were successful; that is, in 2,120 tests the actual author of the sample in question was scored as being the most similar to that sample. Therefore, in terms of raw attribution success, the disputed samples in this experiment were accurately attributed to their author with a success rate of 58.9% (2,120/3,600).

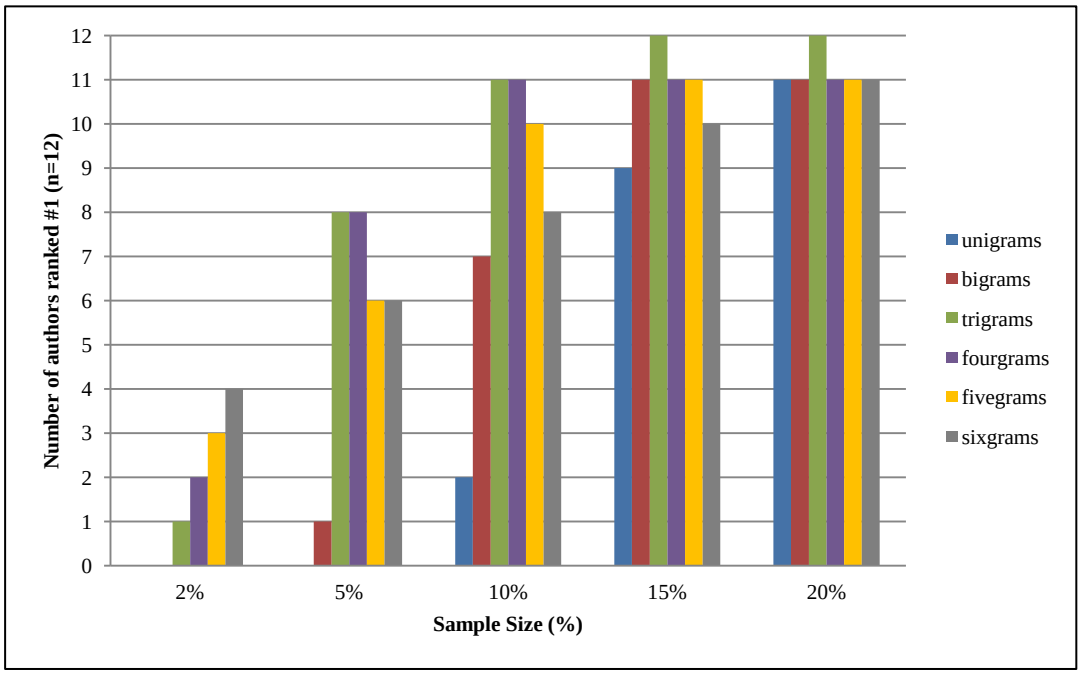
Figure 25. Attribution success rates of all n-gram lengths across all sample sizes. (Grey box highlights all results >70%)



In addition to the attribution of individual samples, mean Jaccard scores were taken to measure how similar the 176 candidate authors were on average to each set of ten samples of each sample size. Allen's 20% samples can be used as an example. Ten different 20% samples were extracted from Allen's set, and each one was tested against the remaining set of his emails, and the full email sets of all of the other 175 authors, using the six different n-gram lengths. In the ten tests using bigrams to compare the sets, Allen achieved the Jaccard scores of 7.69%, 6.72%, 9.79%, 8.75%, 8.62%, 7.43%, 9.36%, 8.94%, 9.75%, 10.02% (the issue of precise Jaccard scores is discussed below in Section 6.5). This gives Allen a mean Jaccard score of 8.71% across his ten samples of this size, which was the highest mean that any of the 176 authors achieved. As such, bigrams have successfully scored Allen as being the most similar, on average, to his 20% samples.

Figure 26 shows the overall results for mean Jaccard scores, the bars representing the number of EEC12 authors (n=12) who achieved the highest mean Jaccard score across their ten samples for each sample size, using all six n-gram types. Overall, the correct author was scored as being most similar on average across their ten samples for each size in 60.83% of cases.

Figure 26. Mean Jaccard results for all n-gram lengths across all sample sizes



Although these success rates are crude, perhaps pessimistic ones, and ones which require discussion in terms of the different sample sizes and different length n-grams, they provide initial answers to the research question of how effective this method is in attributing authorship.

6.3.2 Effect of disputed sample size

In response to the second research question, as is clear by the steep upwards trajectory of lines in Figure 25, the size of the sample being attributed clearly has an effect on the success of the method. Furthermore, as is also clear in Figure 25, the six n-gram lengths produce different success rates with different sample sizes. When attributing the 2% samples (with means of between 89 and 1,317 tokens), the best performing n-gram length is six-grams, achieving an accuracy rate of 37.50%. They are followed by five-grams and four-grams, with success rates of 32.50% and 20.83% respectively. Meanwhile, trigrams achieve only 10.83% accuracy, and unigrams and bigrams perform extremely poorly, successfully attributing less than 1% of the samples. With the 5% samples (means between 253 and 3,367 tokens), unigrams and bigrams still perform very badly, but trigrams improve considerably, achieving a 45.83% success rate. The best performing measure is four-grams, attributing 59.16% of samples, closely followed by five-grams and six-grams both of which have a success rate of 58.33%. By the time the sample sizes are 10% of the

authors' emails (means between 485 and 6,820 tokens), success rates are very high. Again four-grams are the best performers, attributing 86.67% of the samples to their correct authors. This time, they are followed by trigrams with a success rate of 85.83%, while five-grams and six-grams identify the author of the samples correctly 79.17% and 70% respectively, with unigrams and bigrams continuing to underperform. With the 15% sample sizes (means between 712 and 10,116 tokens), accuracy rates become extremely high. Now, trigrams outperform the other measures, achieving a success rate of 92.50%, followed by four-grams at 91.67% and five-grams at 87.50%. This stage is the first in the experiment where bigrams outperform a longer measure, achieving 80% success compared with the 79% of six-grams. Finally, with the largest samples in the experiment of 20% of the authors' emails (means between 951 and 13,436 tokens), all n-gram measures perform extremely accurately. With these large samples, bigrams and trigrams perform equally well at 95% accuracy, which is higher than all the other measures. Four-grams and five-grams achieve 94.17% and 92.60% accuracy respectively, while six-grams return the lowest accuracy of 87.50, being outperformed by individual unigrams at 91.67%.

The Daubert criteria require that the procedures used in the production of admissible expert evidence in US courts should have a known rate of error (Coulthard 2004: 444; Solan and Tiersma 2004: 451). However, there is no indication as to what constitutes an acceptable error rate, and there is not yet a consensus in either the legal or linguistic community as to 'how good is good enough' for a method of authorship attribution. That said, a number of studies (Zheng et al. 2006; Grieve 2007; Koppel et al. 2011) consider 70%–75% accuracy in an attribution task to be 'successful', 'satisfactory', or 'passable'. Following these computational studies, if their 70% threshold is applied then everything above 70% (the shaded box in Figure 25) can be considered 'successful'. This includes the results for four of the measures when attributing 10% samples (trigrams and longer), five measures when attributing 15% samples (bigrams and longer) and all six measures when attributing 20% samples. However, although 70% success may be accurate enough for research purposes, given the high stakes in a forensic context, an error rate of three in ten is clearly not high enough. Instead, those results with accuracy rates of over 90% may be more suitable, which in this study are trigrams

and four-grams when applied to 15% samples and unigrams through to five-grams when attributing 20% samples.

To observe the effect that the size of the sample being attributed has on the results, the number of attributions made for each of the sample sizes by each of the n-gram measures is given in Table 18.

Table 18. Overall success rates across different sample sizes

Sample size/n-gram	Success rate	Sample size/n-gram	Success rate
20%	**667**	**5%**	**269**
unigrams	110	unigrams	0
bigrams	114	bigrams	2
trigrams	114	trigrams	55
four-grams	113	four-grams	72
five-grams	111	five-grams	70
six-grams	105	six-grams	70
Success	**92.64%**	**Success**	**37.36%**
15%	**598**	**2%**	**123**
unigrams	81	unigrams	0
bigrams	96	bigrams	1
trigrams	111	trigrams	13
four-grams	110	four-grams	25
five-grams	105	five-grams	39
six-grams	95	six-grams	45
Success	**83.06%**	**Success**	**17.08%**
10%	**464**		
unigrams	16		
bigrams	62		
trigrams	103		
four-grams	104		
five-grams	95		
six-grams	84		
Success	**64.44%**		

With 20% samples, the overall success rate of this method is very high at 92.64%. That means that of the 720 tests using disputed samples of this size (12 authors, ten random samples, six n-gram lengths), the correct author obtained the highest Jaccard similarity score on 667 occasions (note that, following Juola [2013], the statistical significance of the difference between Jaccard results for different authors was not measured here [Section 4.2]). With the 20% samples, bigrams, trigrams and four-

grams only fail to attribute six or seven samples. All of these unsuccessful samples belonged to one author, Andy Zipper (who is discussed below). Therefore, for the 20% samples of eleven of the twelve EEC12 authors, these n-gram lengths return a 100% accurate rate. Overall success rates for 15% samples are also good at 83.06%. However, the effect of sample size comes into sharp focus as success rates deteriorate, until they fall as low as 17.08% for 2% samples. This is not a surprising result. As Hoover (2001: 423) comments: 'for statistical analysis, the longer the text the better'. Grant (2007) systematically applies lexically-based style markers in discriminating between authors of emails and attributing individual query texts. He found that when using twenty texts per author 81% were classified correctly and all query texts were successfully assigned to their correct authors. However, as the number of texts per author is reduced to three, 'the analysis is seen to break down completely' (Grant 2007: 17). Similarly, Luyckx and Daelemans (2011: 38) use machine learning techniques and a range of lexical-, syntactic- and character-level features and find that 'performance increases with increasing amounts of training data' (Luyckx and Daelemans 2011: 52), and quote Moore (2001) in asserting that 'there is no data like more data'. Eder (2013: 1) performed a similar study in order to 'determine a minimal size of text samples for authorship attribution that would provide stable results'. Using machine learning techniques applied to most frequent words, character n-grams and part-of-speech n-grams on a range of literary corpora of different languages, he found that samples shorter than 5,000 words produced poor results, while for those below 3,000 words 'the obtained results are simply disastrous' (Eder 2013: 4). While all of these studies used different approaches and different types of corpora, they all found that classification and attribution results became less accurate as the amount of data used decreased.

The effect of sample size is also clear in the mean Jaccard results in Figure 26. For the 2% samples, the best results were achieved using six-grams. Four of the twelve authors achieved the highest mean Jaccard score across their ten different 2% samples. In attributing 5% samples, trigrams and four-grams take the lead, scoring as many as eight of the twelve authors as being most similar, on average, to their sets of ten samples. The same two n-gram lengths perform best with the 10% samples, with all but one of the twelve authors (Zipper) having the highest mean Jaccard score across their ten samples of this size. Finally, trigrams continue to improve as the sample sizes increase, and with 15% and 20% samples, all twelve

authors obtained the highest mean Jaccard score across their ten samples of these largest sizes. The importance of these results is that, if a particular individual sample is not successfully attributed, then we can consider how closely the correct author was to being identified. In turn, even if the correct author is not identified for every one of the ten samples, they may be close enough across the ten samples to achieve the highest Jaccard score. Andy Zipper is a good example of this in action. In the results for the 10%, 15% and 20% samples in Figure 26, eleven of the twelve authors achieved the highest mean Jaccard score across their ten extracts using many of the n-gram measures. In each of these cases in which results are accurate for all but one author, the one for which the method does not work is Andy Zipper. However, trigrams *do* score him as being most similar on average to his ten 15% and 20% samples. In terms of his ten individual 15% samples, trigrams only successfully attributed two to him. However, although he did not achieve the highest Jaccard score for eight of the ten samples, he obtained consistently high enough Jaccard score to have a higher mean score than all of the other candidates. The same was the case with his ten 20% samples, only four of which were successfully attributed individually.

This highlights a further possible application of this approach: that of narrowing an initially very large set of candidate authors. Again, Zipper can be used to exemplify this. Zipper does not achieve the highest mean Jaccard score across his 20% sample using unigrams, bigrams, four-grams, five-grams or six-grams. However, even though his mean Jaccard score was not the highest of all 176 authors against which the samples are compared, it does not fall far short. Using unigrams, his Jaccard score (19.61%) was ranked eleventh, with bigrams it was third (5.13%), four-grams it was second (0.38%). When tested with five-grams it slipped back down to tenth (0.14%) and with six-grams it was seventh (0.12%). Therefore, despite not being ranked as the author most similar to his 20% samples using these measures, he was never outside the top eleven. In other words, these n-grams consistently ranked Zipper as being within the top 6% (11/176) of authors. This kind of result indicates that, in addition to accurately identifying authors of individual samples, this word n-gram Jaccard approach may be used as a means of reducing the overall number of candidate authors of a text.

6.3.3 Is size everything?

Although the overall attribution accuracy rate in attributing smaller samples to their correct authors is low in this study, there are two points worth noting. First, as is shown by the results in Table 18 and as becomes clear by following the trajectories of the lines in Figure 25, different n-gram lengths perform better for different sample sizes. More specifically, shorter n-grams perform better than longer ones when attributing larger sample sizes, while longer n-grams outperform shorter ones in attributing smaller samples. In terms of raw attribution, with the 20% samples, the best performing n-grams are bigrams and trigrams. As the samples being attributed shrink to 15% and 10% of the authors' emails, the best performing n-grams become trigrams alone. Finally, with 5% and 2% samples, four-grams and six-grams respectively outperform the others. Similarly, in terms of mean Jaccard scores as presented in Figure 26, although trigrams outperform the other n-grams in the attribution of the 20% and 15% samples, the longer four-grams perform equally as well with 10% and 5% samples. Finally, with the smallest samples of 2%, as with raw attribution scores, six-grams perform best. Therefore, while accuracy generally decreases as the size of the samples being attributed become smaller, it appears as though longer n-grams, particularly four-grams, five-grams and six-grams are more resistant to the reduction in sample sizes than their shorter counterparts. Second, although results are generally poor for 2% samples (only 17.08% accuracy), the method does attribute some very small individual samples to their correct author. A case in point here is the results for Phillip Allen. Allen has the third smallest sample sizes in the experiment and his 2% samples range between 108 and 368 tokens, which is comparable to the very small texts used in some forensic authorship research (e.g. Chaski 2001; 2005; Grant 2007; Rico-Sulayes 2011). The difference here is that there are 176 candidate authors, rather than the much smaller amount used in these other studies. In this experiment, four of these samples are attributed to Allen using trigrams, five using four-grams, seven using five-grams, and six-grams attribute all but one of these ten very small samples to Allen.

Jim Derrick, who has the smallest 2% samples in the experiment is a similar case. His 2% samples comprise between 55 and 145 tokens, and a number of these exceptionally small samples are accurately attributed to him throughout this experiment. Trigrams and six-grams identify him as the author of one of the ten samples, while four-grams and five-grams attribute three of them to him. The

smallest of Derrick's samples to be attributed to him, and the smallest samples to be attributed overall in the experiment, are only 109, 84 and 77 tokens in size. Therefore, although accuracy rates overall decrease with sample size, this method is still successful in attributing exceptionally small segments of data to their correct authors. That said, while the method is accurate in identifying Allen and Derrick as authors of very small samples occurs, it fails to attribute some larger samples to their correct authors. To take five-grams as an example, while successfully attributing seven of Allen's and three of Derrick's very small 2% samples accurately, they only attribute only one of Germany's and Nemec's samples, which are far larger in comparison, ranging between 950 and 1,636 tokens and 793 and 1,447 tokens respectively. Similarly, although Steffes has much larger 2% samples than Allen and Derrick (of between 469 and 854 tokens) five-grams do not successfully identify him as the author of any of these samples. These results indicate that the size of sample being attributed may not be the most important factor in correctly identifying authors. Rather, they suggest that the task of attribution is easier for some authors than others, at least using this method.

Table 19 presents the accuracy rates for each of the EEC12 authors, and these rates further support this suggestion. Although the overall success rate of the method is 58.9% (attributing 2,120 of the 3,600 individual samples), it is more successful than this for most (eight of the twelve authors), and less successful than this for others (four of the twelve). Each of the authors underwent 300 tests (five sample sizes, ten random samples of each size, six n-gram measures). The method was most successful for Lavorato, for whom 242 of his 300 tests were successful. Germany has the most data of the twelve EEC12 authors, and so his sample sizes are consistently the largest, and by some distance (Table 17). However, he does not have the highest success rate, as better results are achieved for Lavorato, Kaminski, Allen and Arnold, all of whom have (much) smaller samples to attribute than him. In spite of having the third smallest samples in terms of tokens, Allen is correctly identified as the author of these samples in 76.33% of cases. Similarly, Derrick's samples consistently contain the fewest tokens at each size, yet the approach performs relatively well in attributing these to him, better than Steffes, Dorland and Haedicke, all of whom have more data.

Table 19. Attribution success rates for individual authors
(mean number of tokens for given sample size in brackets)

	Sample size (total = 60 tests)					Total (n=300)	
Author	**20%**	**15%**	**10%**	**5%**	**2%**		
Lavorato	60 (4,600)	60 (3,360)	54 (2,217)	41 (1,031)	27 (450)	**242**	**80.67%**
Kaminski	60 (9,362)	60 (6,909)	54 (4,656)	40 (2,200)	19 (995)	**233**	**77.67%**
Allen	60 (2,575)	57 (2,001)	49 (1,267)	37 (609)	26 (255)	**229**	**76.33%**
Arnold	60 (4,633)	47 (3,281)	42 (2,481)	37 (1,272)	18 (449)	**204**	**68%**
Germany	60 (13,436)	58 (10.116)	46 (6,820)	29 (3,367)	8 (1,317)	**201**	**67%**
Farmer	60 (4,208)	60 (3,213)	45 (2,016)	21 (1,030)	7 (350)	**193**	**64.33%**
Nemec	60 (9,224)	58 (7,047)	50 (4,963)	22 (2,327)	2 (998)	**192**	**64%**
Derrick	59 (951)	56 (712)	41 (485)	25 (253)	8 (89)	**189**	**63%**
Steffes	60 (6,642)	56 (5,126)	30 (3,033)	9 (1,400)	1 (655)	**156**	**52%**
Dorland	58 (2,464)	42 (1,935)	25 (1,279)	2 (655)	4 (280)	**131**	**43.67%**
Haedicke	56 (3,294)	35 (2,309)	27 (1,591)	5 (780)	3 (286)	**126**	**42%**
Zipper	14 (1,281)	6 (1,090)	1 (554)	1 (259)	0 (151)	**22**	**7.33%**

One author for whom this approach does not perform well with is Andy Zipper. On average, his samples contain fewer emails than anyone else's. However, his mean email length (32 tokens) is far higher than Derrick's (12.9), and so his samples contain more tokens than Derrick's. Nevertheless, for Zipper, the method performs badly, attributing only 7.33% of his samples across the whole experiment. Even for his 20% samples, which range from 787 to 1,778 tokens in size, only 14 of 60 tests (using all six n-grams) were successful. In addition, as was discussed above, in many cases Zipper was the only one of the twelve authors who did not obtain the highest mean Jaccard score across his ten samples for any given size. A possible explanation for this is that although Derrick has more tokens, he has fewer emails than Zipper, and it may be that authors are more likely to repeat word n-grams

across different emails rather than within the same email. Johnson and Wright (2014) conducted a case study of Derrick's authorial style and found that not only were Derrick's emails short, but he was very formulaic and repetitive in his language use across them. For example, twenty three of his emails contained only *Please print the attachment*, nine included *Please format and print the attachment*, and a further seven consisted entirely of *Please handle. Thank you. Jim.* Furthermore, these n-grams were either only found in the emails of Derrick in the Enron corpus, or he was the most frequent user of them (Johnson and Wright 2014: 61). Therefore, Derrick is very consistent, rigid and formulaic in his language use across emails, and that the word n-grams he uses are distinctive to him in the Enron corpus. These factors will have made the attribution of his samples in this study more straightforward than for authors who are more variable, or use less distinctive word n-grams, regardless of the amount of tokens in his samples. At the same time, there is a possible lexical explanation as to why results are so poor for Zipper. As Coulthard (2004: 441) argues, 'the longer a sequence is, the more likely it is that at least some of its components have been created by the open choice principle', as opposed to pre-fabricated, as in the idiom principle (Sinclair 1991). It might be that, in stark contrast to Derrick who seemingly makes frequent and consistent use of formulaic and pre-fabricated language chunks, Zipper is more creative and varied in his collocational choices and lexical phrases. This kind of linguistic behaviour, combined with the relatively small amount of data available for Zipper, may account for why this method struggles to correctly attribute his samples.

The suggestion that the particular author being tested is more important than the size of the samples being attributed is an important one, with implications both for this study and for attribution methods generally. The results here show that word n-grams are more successful in identifying some authors than others. This implies, therefore, that word n-grams are more easily able to capture aspects of some people's idiolects, while for other authors this it is more difficult. In turn, this suggests that while idiolect is manifest in distinctive collocational patterns and lexical sequences for some people, this may not be the case for all people. Theoretically speaking, if we accept that everyone has their own unique idiolect (and the evidence in this and the previous chapter would suggest they do), then it would seem reasonable to also assume that different people's idiolects will be manifest in speaking and writing in different ways. For some, it might be that they

produce the unique collocations and lexical primings that they have developed over their lives in such a way that they can be identified as being idiolectal. For others, it might be that they have distinctive spelling variations of particular words, they may use punctuation in a structurally distinctive way (as found by Chaski 2001; 2005), they may produce distinctive syntactic patterns, or may employ markedly long sentences. Regardless, the important point here is that although the results of this method are good, one linguistic feature, or any finite set of linguistic features, is unlikely to be able to reliably identify idiolectal evidence for every author in any given corpus. As Chaski (2001: 17) argues, despite the promising results she found using measures of syntactically-classified punctuation, it 'should not be used alone in an actual forensic examination, because there is an error-rate' associated with it. This error rate suggests that this particular marker of style was successful with some authors, and less so with others. Stylometric studies have attempted to overcome this problem by combining many different types of features (at lexical, syntactic and character level) in the discrimination and identification of authors. Such studies almost invariably report that features produce better (though not perfect) results when combined than they do independently (e.g. Koppel and Schler 2003: 6; Stamatatos et al. 2001: 212; Grieve 2007: 226; Koppel et al. 2009: 14; Iqbal 2010: 63). As Grant and Baker (2001: 77) note, through such approaches, strange combinations of markers such as the distribution of specific characters and average word length 'might prove to be effective discriminatory components'. However, the fact remains that, unlike the use of collocational patterns and word n-grams, such features lack explicit linguistic and theoretical validity as to *why* they distinguish between people's idiolects.

The fact that different linguistic features may or may not be able to isolate aspects of different individuals' idiolects presents a problem for authorship attribution research. For one, it casts doubt over Butters' (2012: 354) suggestion that forensic linguists should 'establish guidelines on the minimal number and relative strength of the variables that constitute the 'idiolect'. But it also provides a point for future directions for authorship research. An emerging trend in authorship work is to test methods on smaller and smaller datasets, given that the texts available in forensic casework are often very short (Chaski 2001; 2005; Grant 2007; Rico-Sulayes 2011; Luyckx and Daelemans 2011; Eder 2013; Koppel et al. 2013).While this is a necessary trend, future studies would do well to consider the role of the

individuals who constitute the corpora they examine. It is likely that they will find that the features they are employing are more effective in discriminating and identifying some authors than others, irrespective of the amount of data they are represented by. In such cases, it may actually be the individuals who make up the corpus that are influencing the success or failure of the methods, rather than only the amount of data being used.

6.3.4 Which n-gram is best?

The third aim of this chapter was to identify which length of n-gram was most accurate and most reliable in attributing disputed email samples to their correct authors. In this experiment, each of the six n-gram measures was used in 600 tests (ten samples of five sizes for twelve authors). In terms of the number of successful attributions made, four-grams are the most accurate, correctly identifying the author in 424 (70.67%) of the 600 tests they are used in (Table 20).

Table 20. Attribution success rates for the six n-gram measures

	Sample size (total = 120 tests)					**Total (n=600)**	
n-gram	**20%**	**15%**	**10%**	**5%**	**2%**		
four-grams	113	110	104	72	25	**424**	**70.67%**
five-grams	111	105	95	70	39	**420**	**70%**
six-grams	105	95	84	70	45	**399**	**66.50%**
Trigrams	114	111	103	55	13	**396**	**66%**
Bigrams	114	96	62	2	1	**275**	**45.83%**
Unigrams	110	81	16	0	0	**207**	**34.50%**

They are closely followed by five-grams which attribute samples to the correct author in 424 (70%) of the tests in which they were used. There are a total of 60 sets of ten samples across the twelve authors in this experiment, and of these 60, the actual author of the ten samples achieves the highest mean Jaccard score 44 times using trigrams, making this the most effective n-gram measure in terms of mean Jaccard score. They are closely followed by four-grams (43) and five-grams (41). Six-grams follow (39), while bigrams (30) and unigrams (22) are least successful, as they are in terms of the attribution of individual samples. Therefore, the two different ways in which the method has been evaluated here produce different answers to the question: which n-gram is the best? In terms of the

attribution of individual samples it is four-grams, but in terms of mean Jaccard scores it is trigrams. As noted above, authorship studies often report improved results when linguistic features are used in combination to attribute texts compared with when they are used alone. It is possible to combine different word n-grams lengths to measure their success when used together. Because four-grams attribute some samples which trigrams do not (and vice-versa), the results for these two measures combined offer slight improvements to those when they are used independently. When taking the successful attributions of trigrams and four-grams together, 431 of the 600 samples are successfully attributed, an accuracy rate of 71.83%, which is marginally higher than that of four-grams alone. Similarly, if four-grams and five-grams are combined as the best performing n-grams in Table 20, then the authors of 445 of the 600 samples are correctly identified, offering another improved success rate of 74.17%. However, the best combination of two measures is fourgrams and six-grams which together successfully attribute 454 (75.67%) of the 600 samples. Optimum results are obtained when the top four performing n-grams are combined (trigrams to six-grams). When all four of these n-grams lengths are taken together 462 (77%) of the 600 samples are successfully attributed. These results, then, support the argument that combinations of measures outperform the measures used by themselves.

Returning to individual n-gram measures, authorship and plagiarism studies have generally held that the longer a lexical sequence is, the more likely it is to be idiolectal for the speaker or writer using it (Coulthard 2004: 441; Johnson and Woolls 2009: 112; Culwin and Child 2010: 16). Coulthard (2004: 441–2) argues that, with longer sequences, it is less likely that 'the occurrence of this identical sequence in two different texts is a consequence of two speakers/writers coincidentally selecting the same chunk(s) by chance'. Therefore, if the same long lexical string is found in both a disputed text and a known text, this is may be considered strong evidence that these texts have the same author. In turn, it might be expected that the longer n-grams would be more successful than shorter ones in attributing samples in this experiment. On the other hand, however, in the identification of distinctive *deal* n-grams in the emails of Germany and Farmer above (Section 5.3.2), more distinctive trigrams were found than any other length. In fact, there are generally more trigrams and four-grams in a given dataset than any other length. Germany's dataset, the largest in the EEC12, for example, contains

4,803 different unigrams (types), 26,730 bigrams, 41,525 trigrams, 43,087 four-grams, 39,975 five-grams and 35,731 six-grams. The reason for the exponential growth in numbers between unigrams and four-grams is that the same unigram can be included in numerous four-grams. In the sentence *the cat sat on the mat*, the word *on* is found in three different four-grams (i) *the cat sat on*, (ii) *cat sat on the*, (iii) *sat on the mat*. The reason why the numbers plateau at four-grams, as longer n-grams become less common, is that as the sequences extend, they cross sentence boundaries. In this study only those n-grams within sentence boundaries are counted, in the interests of retaining grammatical and semantic information (Section 4.2). Therefore, although longer measures such as six-grams may be more distinctive, authors may be less likely to repeat them. An explanation for the success and accuracy of trigrams and four-grams in this study, therefore, might be that while they are long enough to be distinctive of individual authors, they are also frequent enough to be repeated by authors across emails, so can be found in both disputed samples and their remaining 'known' emails. This supports the hypothesis stated in the previous chapter that word sequences of this length represent units of psycholinguistic encoding that can be used to identify idiolects and attribute disputed material.

However, answering the question of 'which n-gram is best' is not so straightforward. For one, as was discussed above, different n-gram lengths perform better with different sample sizes. Any conclusions are further complicated by the finding that different n-grams perform differently for different authors. Table 21 shows which n-grams accounted for the successful attributions of samples for each of the twelve authors, with the most successful n-gram for each author highlighted in green. As noted above, each author underwent 300 tests (five sample sizes, ten random samples of each size, six n-gram measures). Trigrams, the most accurate n-gram measure according to mean Jaccard results, are the most successful of the six n-grams for only one of the twelve authors, accounting for six of the 22 tests in which Zipper was successfully scored as being the most similar to his samples. Trigrams are also joint best performers with four-grams in attributing Haedicke's samples, both accounting for 22.48% of the 129 tests in which he was correctly identified. Four-grams are also the best performing measures for Farmer, Derrick, Steffes and Dorland, while five-grams attribute more samples correctly to Nemec than any other measures.

Table 21. The performance of the six n-grams across the twelve EEC12 authors (most successful for that author highlighted in green)

	Lavorato (n=242)	**Kaminski (n=233)**	**Allen (n=229)**	**Arnold (n=204)**	**Germany (n=201)**	**Farmer (n=193)**	**Nemec (n=192)**	**Derrick (n=189)**	**Steffes (n=156)**	**Dorland (n=131)**	**Haedicke (n=129)**	**Zipper (n=22)**
Unigrams	25 (10.29%)	24 (10.3%)	19 (8.3%)	12 (5.88%)	18 (8.96%)	20 (10.36%)	20 (10.42%)	22 (11.64%)	21 (13.46%)	14 (10.69%)	12 (9.3%)	1 (4.55)
Bigrams	32 (13.17%)	30 (12.88%)	29 (12.66%)	17 (8.33%)	26 (12.94%)	27 (13.99%)	28 (14.58%)	28 (14.81%)	18 (11.54%)	16 (12.21%)	20 (15.5%)	4 (18.18%)
Trigrams	43 (17.7%)	41 (17.6%)	41 (17.9%)	39 (19.12%)	36 (17.91%)	36 (18.65%)	31 (16.15%)	36 (19.05%)	30 (19.23%)	28 (21.37%)	**29 (22.48%)**	**6 (27.27%)**
Four-grams	45 (18.52%)	44 (18.88%)	44 (19.21%)	43 (21.08%)	40 (19.9%)	**38 (19.69%)**	37 (19.27%)	**39 (20.63%)**	**32 (20.51%)**	**29 (22.14%)**	**29 (22.48%)**	4 (18.18%)
Five-grams	**49 (20.16%)**	**47 (20.17%)**	47 (20.52%)	45 (22.06%)	40 (19.9%)	36 (18.65%)	**39 (20.31%)**	35 (18.52%)	30 (19.23%)	26 (19.85%)	24 (18.6%)	2 (9.09%)
Six-grams	**49 (20.16%)**	**47 (20.17%)**	**49 (21.4%)**	**48 (23.53%)**	**41 (20.4%)**	36 (18.65%)	37 (19.27%)	29 (15.34%)	25 (16.03%)	18 (13.74%)	15 (11.63%)	5 (22.73%)

Five-grams and six-grams perform equally well for Lavorato and Kaminski, the two authors for whom the most samples were successfully attributed. Finally, most of Allen's, Arnold's and Germany's successful tests were performed by six-grams.

It was argued above (Section 6.3.3) that word n-grams are better for capturing elements of some author's idiolects than they are for others. What these results here have revealed is that *particular* word n-grams are most effective for *particular* idiolects. Based on these results, it might be argued that some authors' idiolects, such as Haedicke's and Zipper's, are manifest and are identifiable in shorter collocational sequences than for authors such as Allen, Arnold and Germany. These latter authors, for whom six-grams are the most effective measures in identifying their writing, run counter to the suggestion above that authors do not repeat longer sequences.

Overall, no definitive answer can be given as to which word n-gram is best. Instead, what has been found is something more interesting both in stylistic and idiolectal terms, which also has implications for authorship analysis generally. First, the different n-gram lengths performed differently with different sample sizes. While longer n-grams are more successful in attributing smaller samples, shorter n-grams work better with larger samples. However, it was also found that this method is more successful for some authors than it is for others, in such a way that the amount of data available for the author appears not to be the main influence on the success of the method. Authors with smaller samples sizes, such as Allen and Derrick, had more of their extracted email samples attributed to them than authors with far more data. What remains true is that for this method to work in attributing extracted samples to their actual author, the author's writing needs to exhibit recurring and distinctive collocational patterns and word n-grams. To continue this discussion, Gerald Nemec is used as a case study.

6.4 Attributing Nemec's email samples

Any approach to authorship attribution that can claim to bridge the gap between the stylometric and the stylistic needs to offer linguistic explanations for the statistical results it has produced. This method hinges on authors' uses of recurring and distinctive word n-grams. They need to be recurrent so that they (at least) appear once in the disputed sample email, and (at least) once in the remaining 'known' emails for that author. However, a particular n-gram (e.g. *of the*) may be recurrent in

an author's writing but also recurrent in the writing of all of the other 175 authors. Therefore, the word n-grams also need to be distinctive of the author in question, so that they appear in the disputed samples and *only* the known emails of the author. It is these word n-grams which account for the correct author of samples achieving higher Jaccard scores than others. Once these word n-grams can be isolated in a person's data, they represent a pool of style markers that provide idiolectal evidence for that author, and so can be used to attribute their samples accurately.

Gerald Nemec has been chosen as a case study because, despite having the third largest dataset of the EEC12 authors, the performance of this approach on his data has been mediocre. Only 64% of his 300 tests saw his samples successfully attributed, ranking him as the seventh easiest (or fifth hardest) to identify of the twelve authors, a low rank for the amount of data he has. The samples focused on here will be his 5% samples, ranging between 2,030 and 2,825 tokens, and the n-gram measure examined is five-grams, the most effective length in identifying him as the author of his samples. He is the only one of the twelve authors for which five-grams stand alone as being the most successful measure, suggesting that they capture his style in a way that they do not for the other eleven authors. Five-grams successfully attributed eight of his ten 5% samples. Table 22 presents those five-grams that were found shared between the disputed 5% sample set and the remainder of his emails in at least three of these eight successful tests, as provided by *CFL Jangle* (Section 4.2.). Those in bold in the table are the n-grams that are distinctive of Nemec at population level; they are either not used at all by any of the other 175 authors in the Enron Email Corpus, or they are used by another author, but only once. These are the five-grams that account for Nemec gaining the highest Jaccard score of similarly of all 176 candidates to his eight 5% samples. Also included in the table (not bold) are those which another author does use more than once, but uses far less frequently than Nemec.

The main pattern that emerges from across these distinctive five-grams is that many of them are related to his job as a lawyer in the company and, in particular, are reflective of his collaborative practice with colleagues of drafting and revising legal documents and agreements.

Table 22. Nemec's distinctive five-grams in attributing his 5% samples
(Number in brackets = number of times used by another individual author)

Five-gram	Freq.	Five-gram	Freq.
a clean and redlined copy	**5**	***is a rough draft of***	**7**
a clean and redlined version	**24**	***is the term sheet with***	**3**
a redline with the changes	**3**	*might want to check with*	3(2)
a rough draft of the	**6**	***move to the new garage***	**2**
and redlined version of the	**4**	*please forward to the appropriate*	4(2)
are clean and redlined versions	**5**	***please prepare the form of***	**2**
as an exhibit to the	**2**	***please review and if acceptable***	**11**
attached are clean and redlined	**6**	***please review and lets discuss***	**14**
attached as an exhibit to	**2**	***please review and provide any***	**9**
attached is a clean and	30 (4)	***prepare the form of ca***	**2**
attached is a redline with	**6**	*questions please call me at*	5(2)
attached is a rough draft	**7**	***review and let's discuss***	**11**
attached is the form we	**4**	***the form of ca for***	**2**
attached is the term sheet	**3**	***the term sheet with my***	**3**
be attached as an exhibit	**2**	***to be attached as exhibit***	**5**
clean and redlined version of	**18**	***to make it clear that***	**4**
i am fine with this	5(2)	***to move to the new***	**2**
i am ok with the	5(2)	***with the changes we discussed***	**13**
if you are ok with	4(2)	*you need any further information*	5(3)
is a clean and redlined	**24**	***you please prepare the form***	**2**
is a redline with the	**3**		

Ten of the n-grams in the table include the term *redline* as a noun or *redlined* as an attributive adjective (red in the table). The specialist information technology dictionary *Webopedia* offers the following definition:

> In word processing, *redlining* refers to marking text that has been edited. Typically, redlining is used when two or more people are working on a document together; each individual can *redline* the text he or she has added or edited. The redlined text will then appear in a special color (or as bold) so that others can see the changes that have been made.

This definition matches Nemec's redrafting of documents with his colleagues. All ten of his distinctive five-grams which include this term serve to draw his recipients' attention to a 'redlined' version or draft that he is sending to them. For example:

Example 55

<Subject: ENA-Lost Creek IT Agreement>
<Cc: paul.lucci@enron.com>

Attached is **a clean and redlined copy** of the ENA's IT Agreement on Lost Creek. The revisions incorporate Chris Hoekenga's comments. Please review and let me know if this looks acceptable. If OK we can forward to BR for their final review.

Example 56

<From: gerald.nemec@enron.com>
<To: barry.tycholiz@enron.com>
<Subject: Revised Agency>

Attached is the revised Agency Agreement per our discussion and your notes. **A clean and redlined version** are provided. Please review.

Example 57

<From: gerald.nemec@enron.com>
<To: mark.courtney@enron.com>
<Subject: WT1 LOI>

Attached is **a redline with the changes** we discussed. Please review and if it looks OK, please forward to Susan and Dave.

Example 58

<From: gerald.nemec@enron.com>
<To: stephanie.miller@enron.com, peter.keohane@enron.com>
<Subject: Revised Calpine Docs.>

Stephanie, Attached is a clean **and redlined version of the** Release Agreement and Confirm with the changes we discussed.

Example 59

<From: gerald.nemec@enron.com>
<To: michael.legler@enron.com>
<Subject: Michiwest Agreement>

Attached is **a clean and redlined version of the** IT Agreement. Please review and let me know if you have any questions.

Example 60

<From: gerald.nemec@enron.com>
<To: greg.brazaitis@enron.com>
<Subject: Pace Interconnects>

Attached are clean and redlined versions of the interconnect agreements with revisions. Please review and if okay foward to Pace for their review.

As part of his job, Nemec finds himself in the communicative situation in which he is required to forward on edited or 'redlined' versions or copies of documents to his colleagues for their review, as exemplified in Examples 55–60. This recurrent and familiar situation has subsequently given rise to a number of potentially formulaic and pre-fabricated collocational sequences, which Nemec employs to fulfil the purpose of his email. As well as being recurrently used by Nemec, they are all also distinctive of his authorial style, as they do not appear in any of the emails of the other 175 Enron employees. Given this population-level distinctiveness, they provide evidence for Nemec's idiolect, when compared with the population and linguistic community from which he is taken. This underlines the difficulty of divorcing content words from the notion of idiolect, and highlights how such words can and should be central to the analysis of individual linguistic variation. Based on the evidence here, these five-grams are idiolectal for Nemec, and the means by which they have entered and have been retained in his idiolect is through his occupation. Although they may have become part of his idiolect through his occupation, others with the same occupation do not use them. *Redline(d)* is a specialist term, and one Nemec uses frequently (n=150), but he is not the only author to use it. Kay Mann (n=35), Sara Shackleton (26) and Debra Perlingiere (23), all of whom are also lawyers, also use this term frequently, albeit not as frequently as Nemec. Despite these authors' shared use of this term, the way in which Nemec uses it, and the way in which he expresses that he is forwarding a redlined version of a document, is unique. Instead, Shackleton uses phrases such as *attached is my redline*, Perlingerie writes *here is a redline and clean version*, and Mann uses *here's the current redline.*

The case is similar for Nemec's distinctive five-grams that include the words *draft, term sheet* and *exhibit* (blue in the table). As with *redline(d),* Nemec's distinctive n-grams involve him expressing to his recipient that he has attached documents for their review:

Example 61

<From: gerald.nemec@enron.com>
<To: don.baldridge@enron.com>
<Subject: Conoco Agreement>

Attached is a rough draft of the agreement we discussed last week. Please review and let's discuss.

Example 62

<From: gerald.nemec@enron.com>
<To: mark.whitt@enron.com, theresa.staab@enron.com>
<Subject: Purchase Supplement Letter>
Attached **is a rough draft of** the pricing letter for Kennedy Oil on the excess gas volumes. Please review and add the relevant pricing structure. Let me know if you have any questions.

Example 63

<From: gerald.nemec@enron.com>
<To: rusty.belflower@enron.com>
<Subject: Master Agreement for E&I Work>

Rusty, Attached is a blank form for use in the E&I work. The Work Offer form for the other contracts should be **attached as an Exhibit to** the other contracts. If it is not, let me know.

Example 64

<From: gerald.nemec@enron.com>
<To: miguel.vasquez@enron.com>
<Subject: Falcon Term Sheet>

Attached is the term sheet with my revisions. Please review.

Again, these are terms that Nemec shares with other Enron employees, but encodes in unique collocational sequences. He uses *draft* 129 times, which is fewer than Perlingiere (207), Shackleton (176), and Mann (166), and slightly more than Tana Jones (96), also a lawyer. However, Nemec is only one to premodify *draft* with *rough*, and his subsequent sequence *attached is a rough draft of the* includes three distinctive five-grams (*attached is a rough draft, is a rough draft of* and *a rough draft of the*). Similarly, Nemec uses *term sheet* twelve times, and it also appears in Mann's (19), Ward's (11) and Shackleton's (6) emails, while *exhibit*, which he uses 35 times, is also used frequently by Mann (22), Shackleton (22), Perlingerie (9) and Susan Scott (9). Despite this, Nemec's emails contain distinctive five-grams which include these terms.

Another five-gram that is distinctive of Nemec is *with the changes we discussed*. As he forwards on the edited documents to his recipients, he makes explicit reference to previous communication they have had with regard to these documents, in such a way that indicates the intertextual nature of the documents and their revision process:

Example 65

<From: gerald.nemec@enron.com>
<To: barry.tycholiz@enron.com>
<Subject: Nevada Power Confirms>

Barry, Attached are both confirms for the Nevada Power master **with the changes we discussed** yesterday. Please review and if acceptable, please forward Nevada Power.

Example 66

<From: gerald.nemec@enron.com>
<To: barry.tycholiz@enron.com>
<Subject: Revised Docs.>

Barry, Attached are redlines of the docs **with the changes we discussed** today. Please pay special note to the Maximum Daily Deliverability Quantity of the Confirm. I want to discuss how this limits us and what flexibility this section needs. Please review and lets discuss.

This is a sequence which Nemec uses thirteen times, suggesting that it is a feature he routinely and habitually uses to refer to the documents he is forwarding, and acknowledging that he has addressed the issues raised in previous correspondence with his recipients. In addition, Example 66 ends with another five-gram that Nemec repeatedly employs: *please review and lets discuss*. He uses this a total of 25 times, eleven times with an apostrophe in *let's* and fourteen times without, both of which are only found in his emails in the Enron Email Corpus:

Example 67

<From: gerald.nemec@enron.com>
<To: david.marshall@enron.com>
<Subject: FW: LRCI, Inc. 113A>

David, Attached is the document with the insurance mark-ups that I noted on my voice mail. **Please review and lets discuss.**

Example 68

<From: gerald.nemec@enron.com>
<To: staci.holtzman@enron.com>
<Subject: Revised CSA>

<Cc: mark.knippa@enron.com, chris.hilgert@enron.com>
<Bcc: mark.knippa@enron.com, chris.hilgert@enron.com>
Staci, Attached is the CSA with my comments redlined. **Please review and let's discuss.**

Whereas Nemec's distinctive use of *with the changes we discussed* refers to the textual history of the document in question, his distinctive *please review and let's discuss* refers to the textual future of the document. It serves to indicate that the participants in this communication will engage with each other again, after this email, in the collaborative drafting of this document. A similar function is fulfilled by Nemec's distinctive five-grams *please review and provide any* and *please review and if acceptable*:

Example 69

<From: gerald.nemec@enron.com>
<To: mark.whitt@enron.com, steve.pruett@enron.com, scott.josey@enron.com>
<Subject:>
<Cc: barbara.gray@enron.com, audrey.o'neil@enron.com>

Attached is a draft of a nonbinding letter for use to submit the indicative bid to Wildhorse. **Please review and provide any comments.**

Example 70

<From: gerald.nemec@enron.com>
<To: greg.brazaitis@enron.com>
<Subject: Pace Interconnects>
<Cc: eric.gillaspie@enron.com>

Attached is a clean and redlined versions of the Gateway and Rio Nogales with the revisions we discussed. **Please review and if acceptable**, please forward to Pace. Merry Christmas.

Nemec is not the only author to politely direct his recipient to review documents. He is the most frequent user of *please review*, writing it 254 times across his email dataset. However, besides him, it appears 264 times in the rest of the corpus, and is used by an additional 58 authors. What begins to distinguish Nemec from his colleagues is the use of *and* after *please review*. *Please review and* is used 124 times by Nemec, and the author who uses it next most frequently is Dan Hyvl with eleven instances. By the time the string becomes five words long, *please review and provide any* and *please review and if acceptable* are both unique to Nemec in the corpus. Therefore, in the same way as above with *redline(d), draft* and *term sheet*, the directive *please review* is a shared element of communicative practice across Enron employees. Yet despite this commonality and shared linguistic repertoire, author-distinctive linguistic preferences and patterns can still emerge.

In theoretical terms, the five-grams discussed here offer a snapshot of Nemec's idiolect. Despite sharing specialist terms and collaborative practices with his Enron colleagues, he has developed unique linguistic means by which to fulfil recurring communicative functions. They serve to complement the analysis in Chapter 5 in that they further demonstrate how content words (*redline, draft, term sheet, exhibit*) can be central to discussions of idiolectal variation, in that they are linguistically encoded and collocationally packaged differently by different writers. Furthermore, they support the argument that people express the same speech act (*please review*) in distinctive ways. In methodological terms, these word n-grams offer linguistic explanations for the statistical results obtained in the experiment. By examining the features that were responsible for the successful attributions of his samples, and testing their population-level distinctiveness, it is possible to isolate some of the most important features in identifying Nemec's writing.

6.5 Evaluating the word n-gram approach

The word n-gram Jaccard approach developed here has produced some very accurate results in the attribution of disputed email samples, achieving success rates as high as 100% for larger samples sizes of some authors. However, a few methodological caveats are worthy of note.

Cheng (2013: 547) expresses his concern that although complex statistical models and machine learning algorithms in authorship research produce good results, they may confuse or mislead the lay jury. Jaccard's similarity co-efficient is essentially a percentage. It takes into account all the number of word n-grams of any given length shared across two datasets, as a proportion of all of the word n-grams of that length in the two datasets combined (see Section 4.4.3). The simplicity of this measure perhaps appeals to the mathematical abilities that can be expected of a lay jury more so than more complex algorithms do. However, given the proportional nature of Jaccard, the scores produced for the longer n-grams can be very small.

Table 23 shows how the Jaccard score of similarity between an extracted sample and a comparison file changes as the n-gram measure used gets longer.

Table 23. The effect of n-gram length on Jaccard score

n-gram	Filepair	Shared	Unique to Sample	Unique to Comparison File	Combined	Jaccard
unigrams	germany{80}	1681	540	2582	4803	35.00%
bigrams	germany{80}	3772	4169	18787	26728	4.11%
trigrams	germany{80}	2423	7425	31676	41524	5.84%
four-grams	germany{80}	1228	8151	33707	43086	2.85%
five-grams	germany{80}	759	7702	31521	39982	1.90%
six-grams	germany{80}	558	6923	28250	35731	1.56%

These are the results of the same 20% sample of Germany's emails when tested using all six n-gram measures. In measuring the raw attribution success of this method, all that was taken into account was which of the 176 authors achieved the highest Jaccard score when compared with any one extracted sample. As the results in the 'Filepair' column in Table 23 show, Germany's remaining 80% samples were consistently measured as being most similar to this 20% sample. That is, Germany was successfully identified as the author of this sample using all six n-gram measures. However, what can be observed is that while Germany's remaining 80% samples achieve a Jaccard score which is 35% similar to his sample using unigrams, this similarity drops to 1.56% for six-grams. The Jaccard score is calculated by dividing the value in the 'Shared' column with the value in the 'Combined' column and multiplying by 100. Generally speaking, as the length of the n-gram increases, so too does the number of n-grams in the dataset. As such, the 'Combined' value increases. At the same time, the number of items shared between the sample and the comparison texts decreases, as they are repeated less and less by authors. As a result, the Jaccard scores decrease. While being easily explainable, and inherent in this kind of proportional statistic, such low similarity scores as 1.56% between sample and comparison text may seem too low to non-experts to be a reliable attribution of authorship. However, what must be stressed here is that, although low, these Jaccard scores were the highest that any of the authors in the corpus attained in this test, and that is the basis of the attribution.

A second related caveat is that, because Jaccard measures similarity between the sample set and the 'known' comparison file, one comparison file is always going to be ranked as most similar. When this file is the remaining emails of the author

who is responsible for the sample, this is a successful attribution. However, when the highest ranked comparison file is of someone else, the disputed samples are effectively being misattributed. In order to overcome this, and minimise the effect of misattributions, the method has been tested on ten different random samples of each size and for each author, and mean Jaccard scores across the ten different samples have been calculated. Nevertheless, for the 2% samples, for which the method successfully identified the author 17.08% of the time, there is a misattribution rate as high as 82.92%. Thankfully, the misattribution rates for 20% and 15% samples were as low as 7.36% and 16.94%. In this experiment, because the actual authors of the texts are known, it is straightforward to identify misattributions. However, in a forensic context in which the actual author of a text is not known, it is impossible to spot misattributions, and the cost of such errors is far higher. In an attempt to avoid erroneous misattributions, Grant (2007: 20), in developing his quantitative approach to authorship attribution, distinguishes between tests in which the questioned document was correctly classified, misclassified and where no classification was possible. He (Grant 2007: 22) argues that methods of authorship analysis need 'to be able to describe the line where attribution becomes impossible' and for his study, he uses a threshold which already exists for methods using log-likelihood statistics (Champod and Evett 2000; Rose 2002). The development of such a threshold for Jaccard scores could identify a particular level of similarity below which it would be concluded that no reliable attribution could be made. This would provide a more nuanced set of results as opposed to a binary correct/incorrect attribution, and would reduce the number of misattributions. Investigating whether such a threshold can be applied to Jaccard scores, and what form it might take, is a matter for future study.

A final methodological comment worth noting relates to dataset size. In this experiment, the size of the 'disputed' sample set being attributed has been carefully controlled, ranging from between 20% and 2% of authors' emails. In contrast, the size of the datasets against which these samples are being compared has not been controlled. This means that each sample in this experiment is being measured against the full datasets of the other 175 authors in the corpus, which range from 170,316 tokens for Jeff Dasovich and 20 tokens for Gretel Smith. This decision to use all of the data for each of the candidate authors was a deliberate one, given that in forensic casework, the analyst is likely to wish to use all of the data at their disposal in reaching a judgment about authorship. For empirical purposes, however,

future research testing this method will control the size of the comparison, or 'known', sets as well as the sample sets, in order to investigate how this affects the accuracy of the method.

As McMenamin (2002: 166) notes, 'the process of testing the theory and practice of authorship identification is continuous', and Chaski (2001: 41) highlights the importance of the replication and re-testing of proposed methods. Indeed, further empirical testing, using both the Enron corpus and alternative datasets, is required to assess the word n-gram and Jaccard approach and to address the caveats outlined here. Such research would serve to evaluate the effectiveness and reliability of the method and, ultimately, the extent to which it is ready for reliable use by forensic linguists.

6.6 Chapter conclusion: bridging the gap

The aim of this chapter was to develop a methodology for authorship analysis that bridges the gap between stylometric and stylistic approaches. The aim was to combine the objectivity and statistical reliability of stylometric techniques with the linguistically-motivated and theoretically-underpinned analyses of stylistic approaches.

The method subsequently developed and tested is one which uses word n-grams and Jaccard's co-efficient to measure similarity between extracted 'disputed' email samples and the remaining 'known' email sets of 176 Enron employees, who collectively represent a large pool of candidate authors. This method, it is argued, can bridge the gap between stylometric and stylistic approaches to authorship analysis, and combine the best aspects of the two. First, before any attributions were attempted, word n-grams were selected as the linguistic features on which the method is based, given that the analysis in Chapter 5 provided evidence in support of the theory that collocations and lexical co-selections were idiolectal for individual writers. Therefore, the selection of word n-grams as features for comparison has theoretical underpinning. Second, the actual attribution task was performed using Jaccard's co-efficient as a statistical measure of similarity, and was systematically and rigorously tested using ten random samples, of five different sizes, on twelve different authors. This produced a set of success and error rates for each of the six n-gram lengths, on each author, and each sample size. Finally, using a case study of one author—the lawyer Gerald Nemec—the statistical results were supported by

linguistic results. It was possible to identify a specific pool of five-grams that were responsible for the successful attribution of his 5% email samples. These n-grams are recurrent across his emails and distinctive of him at population level in the Enron corpus, and as such can be considered as being elements of his unique idiolect. In particular, most of these lexical sequences represented distinctive collocational preferences for Nemec in relation to everyday communicative elements and activities in his role as a lawyer. Reinforcing the strength of this idiolectal evidence is the fact that these distinctive collocation patterns were found within a linguistic community (both of lawyers and Enron employees generally), who share a common linguistic code.

The discussion of the results of this chapter were focused on three research questions, which can be summarised as: (i) does the method work? (ii) what effect does reducing the size of the sample being attributed have on the accuracy of the method? and (iii) which of the six n-gram measures is most effective? Given the very small sizes of some of the samples of the EEC12 authors used in the experiment, and the high number of 176 candidate authors for each sample, the attribution task in this experiment was a relatively difficult one. It became clear quickly in the discussion of the results that none of these three research questions had straightforward answers. First, the method performed with an overall crude success rate of 58.9% across the 3,600 individual tests. In terms of mean Jaccard scores, the correct authors were scored as being most similar on average across their ten samples for each size with a success rate of 60.83%. However, the success of the method was clearly influenced by the size of the samples being attributed. When attributing 20% samples, for example, an average total success rate of 92.64% was achieved across all the six n-gram measures. Within this, however, bigrams, trigrams and four-grams had a 100% success rate for eleven of the twelve authors tested. As the sample sizes decreased, so too did the accuracy rates, until only a 17.08% success rate was achieved with the 2% samples. That said, the longer n-grams of between trigrams and six-grams did correctly identify Derrick as the author of samples as small as 109, 84 and 77 tokens in size. What was also discovered is that the method worked better with some authors than others, apparently independent of the amount of data the authors have. For instance, 80.67% of Lavorato's tests were successful, compared with only 7.33% of Zipper's. This brought into question whether the size of the samples, or the author of the samples,

was the most important factor in the success or failure of this or of any other method of authorship attribution. Finally, there is no straightforward answer to which n-gram is the most accurate and reliable for attributing authorship of Enron emails. In terms of raw attribution successes four-grams were the most effective, but in terms of mean Jaccard scores trigrams prevailed. Further complicating the results was the finding that certain n-grams performed better with certain sample sizes. One overall observable pattern was that longer n-grams outperformed shorter ones with small samples sizes, while the reverse is the case for larger samples. Moreover, the same n-gram length was not the most accurate across all authors. Rather, certain n-grams were found to be better suited to identifying the writing of certain authors. This final point indicates that while some authors' idiolects may be manifest in longer lexical sequences, others' may be identifiable in much shorter ones.

In the same way as this method aims to combine stylometric and stylistic approaches to authorship analysis, other recent studies have also pursued this goal (Section 2.2.4). One approach which has received attention from both linguistic and computational camps is the use of Systemic Functional Linguistics (SFL) (Halliday and Matthiessen 2004) as a theory of language which can explain linguistic variation and therefore be used in authorship analysis (Argamon and Koppel 2010; 2013; Nini and Grant 2013). In the same way as this study has argued that authors' idiolects are constructed through their unique linguistic experience, these studies hold that what shapes the idiolect or 'code' of an individual is 'the experience of a certain context that the individual has gathered, which is in turn shaped by their social background' (Nini and Grant 2013: 180). However, in contrast to the specific focus on collocational preferences and distinctive word n-grams proposed over the last two chapters, these approaches use SFL as a framework that 'describes every level of language within one single model, starting from phonology to pragmatics up to sociolinguistic variation' (Nini and Grant 2013: 178). On the one hand, applying such a holistic approach to analysing variation at every level of the linguistic system goes some way to addressing the issue raised above (Section 6.3.3) that different authors' idiolects are manifest in different ways. In the same way as in the present study, Nini and Grant (2013: 185, 188) are successful in identifying elements of 'personal codal variation' in a situation where there is a good deal of homogeneity across the texts and authors that they examine. On the other hand, Nini and Grant (2013: 188–9) identify a number of practical methodological issues with the SFL

approach. Not only does it involve a significant expenditure of time and effort in coding data in such fine detail, but the coding is performed manually and so is susceptible to subjective (and therefore unreliable) analysis. These issues are not apparent in the word n-gram approach, as the identification of extremely large numbers of discrete word clusters is a straightforward and objective task for a computer. Similarly, it might be argued that by considering collocation patterns and lexical co-selections, this approach is capturing, by proxy, many of the detailed aspects of lexical and grammatical variation coded for within SFL. This was demonstrated in the analysis in Chapter 5 which identified variation in both lexical and grammatical preferences across authors. Finally, such highly detailed coding in the SFL approach produced too many variables to feed into a useful discriminatory statistical model (Nini and Grant 2013: 184). In comparison, the Jaccard statistic as a measure of similarity can handle as many word n-grams (or other linguistic features) as required.

The comparison of the approach developed in this chapter and that of SFL is not to claim that one is more useful or reliable for authorship analysis than the other. Argamon and Koppel (2013: 302) make clear that: ‘we do not claim, of course, that SFL is the only, or even necessarily the best, approach’. This sentiment is echoed here with regard to the word n-gram measure. Both approaches represent a welcome shift in research, working together towards bridging the gap between stylometric and stylistic methodologies in authorship analysis.

The third and final analysis chapter, which follows, aims to extend this combined stylometric and stylistic approach to the problem of author profiling. At present this field is dominated exclusively by quantitative approaches and statistical evidence.

7 Author profiling of Enron employees

Author profiling presents a different kind of task to that of author attribution. Rather than attributing disputed texts to a particular author, author profiling seeks to determine the social characteristics of a text's author (Argamon et al. 2013: 307). In Section 2.4.1 parallels were drawn between author profiling in a forensic linguistic context, and offender or criminal profiling in a forensic psychology context. Within the latter, there are two established methodological approaches: 'nomothetic' approaches, which search for general quantitative patterns based on the combined data from many individuals, and 'idiographic' approaches, which are the qualitative and intensive studies of single individuals. Author profiling is a relatively young field, and is beginning to develop at a time when stylometric approaches to authorship *attribution* are extremely common and increasingly sophisticated. Therefore, it is no surprise that all of the existing research into authorship *profiling* is stylometric—or nomothetic—in nature, statistically correlating linguistic features with social variables such as age, gender, native language, level of education and personality type (e.g. Argamon et al. 2003; 2009; 2013; Noecker et al. 2013; Pham et al, 2008; Luyckx and Daelemans 2008; Estival et al. 2007). There are criticisms of such quantitative nomothetic methods in both author and offender profiling. Coulthard et al. (2011: 538) point out that in a linguistic context, generalisations based on large groups of language users may not be applicable to any one individual. Similarly, in a criminal profiling context, Turvey (2012: 122) argues that nomothetic averages cannot be relied upon as they do not describe real offenders. In contrast, he argues that idiographic psychological profiles are more 'concrete', as they describe actual offenders who exist in the real world.

The dichotomy that exists between nomothetic and idiographic approaches to criminal profiling is akin to the divergence of stylometric and stylistic methodologies in authorship attribution. The previous two chapters have worked towards developing an approach which combines the two. In the same vein, this chapter sets out to contribute to the practice of authorship profiling by developing and combining nomothetic (quantitative) and idiographic (qualitative, case study) approaches in a forensic linguistic context. Using the almost 2.5 million word Enron

Email Corpus 80-author sample (EEC80) created especially for authorship profiling analysis (Section 3.4.4), this chapter has two primary research aims:

1. To use a nomothetic statistical approach to identify which linguistic features can be used to distinguish between different groups of authors, particularly between males and females and employees with different occupational roles in the Enron corporation.

2. To use a more idiographic stylistic approach in analysing how such potentially discriminatory features are used in context, both across and within these social groups, and identify how the use of these features reflects various aspects of authors' identities.

To achieve the first aim, a quantitative approach is used to identify features for which there is a statistically significant difference between how frequently they are used by authors from either gender and across eight occupation types in Enron: (i) Presidents, CEOs and COOs, (ii) Vice Presidents, (iii) directors and managing directors, (iv) lawyers, (v) managers, (vi) traders, (vii) analysts, specialists and associates, (viii) assistants. The second aim moves beyond the quantitative generalised nomothetic results, and adopts a more qualitative analysis of the different ways in which such discriminatory features are used stylistically (e.g. in collocations and word n-grams) and contextually (e.g. with different recipients and with different purposes) by social groups and individual authors within these groups. The argument that is advanced in this chapter is that author profiling, which aims to correlate quantifiable linguistic features with discrete social categories, relies on a one-dimensional conceptualisation of author identity. It is hypothesised here that authors and groups of authors make language choices in response to different contextual demands, such as the purpose of the communication and the relationship that they have with the recipient, in such a way that they draw upon and project particular aspects of their identity, rather than because they belong to one social category or another.

The implication of this for forensic author profiling is that the established notion of an *author profile*—the age, gender, personality type, occupation, native language—can be supplemented by an understanding of patterns of linguistic and communicative behaviour of individuals, moving us towards a concept of an author's *linguistic profile.* Combined, such approaches offer richer more linguistically explainable evidence, informed by a nuanced appreciation of author

identity, to aid in forensic investigations and help to answer the question ‘what kind of linguistic person wrote this text?’ (Grant 2008: 223).

7.1 Discriminating between genders and occupations

This first section of analysis identifies overall general linguistic differences between the 46 male and 34 female employees in the EEC80 sample, and between the eight different occupational categories. The linguistic variables that are used in this nomothetic analysis follow the tradition set out by existing computational approaches to author profiling: function words and parts-of-speech. These two linguistic elements reflect the distinction that Argamon et al. (2009: 119) draw between ‘style’ and ‘content’ based features. Function words (e.g. articles, determiners, pronouns) are ‘style’ based features in that they are more independent of topic, while ‘content’ based features (e.g. nouns, verbs, adjectives, and adverbs) vary across topics. While the usefulness of individual function words in authorship attribution is well documented (see Section 2.2.2), their value is still to be fully evaluated in author profiling contexts. In this analysis, whereas the frequency of individual function words (e.g. *the, and, for*) is counted, the frequency of individual content words is not. Instead, only the overall cumulative totals of such words are counted, so that it will be possible to compare whether men use more adjectives overall than women, for example, but not whether they use *happy, fun* or *good* more frequently. The motivation for this decision is partly that if content words were considered individually, then the pool of features used would be too large (with many features occurring with very low frequency) to feasibly discuss here. However, the main motivation is that previous research (e.g. Thompson and Murachver 2001; Koppel et al. 2002; Boulis and Ostendorf 2005) has often based comparisons of gender on content words, finding, for example, that men use words such as *software, democracy, dude* and *shit* more frequently than women, who prefer words such as *cute*, *boyfriend*, *pink* and *mommy*. The issue with such findings is that these words are so heavily dependent on topic that the distinctions being observed in these studies is more likely to be overall cultural differences between males and females (Coates 2004; Tannen 1990) rather than pure linguistic differences. Furthermore, not counting individual content words avoids reporting obvious unremarkable differences across occupation groups such as traders using *deal, buy* and *sell* more than lawyers (which is the case). By using function words

and parts-of-speech, this analysis bypasses describing differences in the topics about which different groups are emailing, and focuses on identifying stylistic differences between the groups. Using the CLAWS tagger (see Section 4.3.1), the full EEC80 corpus was tagged for a total of 291 features: 270 individual function words, eleven function word POS categories and ten content POS categories (Table 24). As well as individual function words, superordinate function word POS categories were also counted.

Table 24. Linguistic features used in the profiling of EEC80 authors

Function words	**Content POS categories**
articles (3) (*the, a, an*)	adjective (exc. comp. & superl.)
determiners (35) (e.g. *this, these, those, all, some, most*)	comparative adjectives
personal pronouns (17) (e.g. *I, me, you, she, he, it*)	superlative adjectives
indefinite pronouns (14) (e.g. *anyone, everything*)	adverbs
reflexive pronouns (10) (e.g. *myself, himself, herself)*	singular nouns
total interjections (64) (e.g. *aah, hi, mmm, oh, whoa*)	plural nouns
conjunctions (27) (e.g. *and, but, because, or*)	base form verbs
modal verbs (9) (e.g. *can, could, must, will, would*)	past tense verbs
prepositions (70) (e.g. *at, in, for, on, with*)	*-ing* verbs
wh- words (18) (e.g. *what, where, why, who*)	*-s* verbs
please/thank(s) (3) (*please, thank, thanks*)	
Total= 281 (11 POS categories and 270 individual words)	Total= 10

7.1.1 Sex differences

Although the label ‘sex’ was used when categorising Enron employees in the building the EEC80 sample (Section 3.4.4), as the discussion now shifts to the language use of the individuals who make up these groups, the term ‘gender’ is preferred here. It is well established in sociolinguistics that gender is a social construction of the binary categories of sex (Eckert 1989: 253), and is constructed through a complex array of social practices of which language is one (Eckert and McConnell-Ginet 1992: 484). Therefore, ‘gender’ will be used throughout this analysis. Furthermore, ‘gender’ is the preferred term in existing author profiling research (e.g. Argamon et al. 2003; Bamman et al. 2014).

The Mann-Whitney U tests (see Section 4.4.2) were used to compare the frequency with which male and female authors in the EEC80 use the linguistic features. The results find that for 35 of the 291 features there is a statistically

significant difference in male and female use in the EEC80 corpus: three function POS categories (total articles, reflexive pronouns and *wh*- words), two content POS categories (total comparative and superlative adjectives) and thirty individual function words (Table 25). The first thing to note is how few of the 291 features actually differentiate between the genders, only 12.03% (35/291). With the vast majority of features (87.97%) there is no significant difference in the frequency with which they are used by men and women. For the most part, then, the email writing of males is indistinguishable from that of females in the EEC80 sample. This was also the case in Argamon et al.'s (2003) study, which began with over 1,000 topic-independent features, fewer than 50 of which (5%) were found to be useful in distinguishing male-authored texts from female-authored texts. Even in studies that draw on content features, there are often very few items that significantly differentiate between male and female writers (e.g. Thompson and Murachver 2001: 199). This apparent lack of difference in the language of males and females has also emerged from sociolinguistics more generally. Baker (2014: 19) notes that 'within academic research, it is increasingly common to read criticisms of the "gender differences" paradigm'. He cites Hyde's (2005) meta-analysis survey of dozens of studies of verbal and behavioural gender differences, which concluded that most studies found that the overall difference between men and women was either very small or close to zero. Cameron (2008: 3) goes as far as to argue that 'the idea that men and women differ fundamentally in the way they use language to communicate is a myth'. Johnson (1997: 11) suggests that linguists have been so preoccupied with uncovering statistically significant differences between men and women that they 'frequently seem to overlook one important fact: the two sexes are still drawing on the same linguistic resources'. A preoccupation with difference at the expense of exploring similarity is a vice of comparative corpus linguistics generally (Taylor 2013), but author profiling has a vested interest in emphasising significant differences between groups of authors, not only for forensic applications, but also for commercial and marketing purposes (Argamon et al. 2009: 119).

Table 25.Thirty-five features for which there is a statistically significant difference between male and female authors (Mann-Whitney U, p<.05)

	Mean Rank		**Sig.**	**M/F**
	Male	**Female**	**(p=)**	**feature**
total articles	46.63	32.21	.006	Male
total reflexive pronoun*	47.46	31.09	.002	Male
total *wh-* adverbs*	45.4	33.87	.028	Male
total comparative adj.*	46.97	31.75	.004	Male
total superlative adj.	47.84	30.57	.001	Male
*and**	45.23	34.1	.034	Male
another	46	33.06	.013	Male
*around**	46.87	31.88	.004	Male
*down**	47.89	30.5	.001	Male
fewer	42.35	38	.048	Male
for	34.95	48.01	.013	Female
half	44.72	34.79	.040	Male
hi	34.79	48.22	.005	Female
huh	44.01	35.75	.011	Male
*I**	46.59	32.26	.006	Male
*into**	45.04	34.35	.041	Male
less	48.37	29.85	.000	Male
*more**	47.01	31.69	.004	Male
most	44.91	34.53	.045	Male
much	45.65	33.53	.020	Male
*my**	46	33.06	.014	Male
ought	42.72	37.5	.030	Male
*ourselves**	44.46	35.15	.006	Male
*over**	46.77	32.01	.005	Male
*some**	47.46	31.09	.002	Male
thanks	34.21	49.01	.005	Female
the	48.46	29.74	.000	Male
toward	44.64	34.85	.005	Male
*what**	45.95	33.13	.015	Male
*where**	46.91	31.82	.004	Male
wherever	42.35	38	.049	Male
*why**	44.92	34.51	.046	Male
yep	43.35	36.5	.011	Male
*yo**	42.35	38	.049	Male
yourself	46.88	31.87	.000	Male

**also distinguish between occupations*

(N.B. Male and female mean rank combined does not equal 100)

The suggestion that linguistic differences between men and women are overstated immediately problematises the process of nomothetic author profiling. So too does the fact that the results here from EEC80 do not correspond with those of other gender-focused author profiling studies. In Table 25 all but three of the features identified are predominantly male features. Within these, male authors use more articles (*the, a, an*) than females, particularly the definite article. This corresponds with what was found by Heylighen and Dewaele (2002: 303), Koppel et al. (2002: 408) and Argamon et al. (2003: 325), but conflicts with Bamman et al. (2014: 142) who found no significant association between article use and either male or female writing. The same is the case for the determiners *another, less, more, most, much, my* and *some,* which are all used significantly more by male employees in EEC80. This is not simply a result of men making a greater use of common nouns, as no significant difference was found between genders for those. This supports what Argamon et al. (2003: 334) and Koppel et al. (2002) found, and have accounted for by the argument that 'male authors are more likely to 'indicate' or 'specify' the things that they write about' and that they 'reliably provide more specification' (Argamon et al. 2003: 334). However, Herring and Paolillo (2006: 449) found that demonstratives were more common in female blog writing, while Burger et al. (2011: 1305) found *my* to be a female feature, the opposite of what was found in EEC80. The results for the EEC80 find that the prepositions *around, down, into* and *over* are all male features. Heylighen and Dewaele (2002: 303) and Argamon (2003: 334) also found prepositions to be more common in the writing of males. Bamman et al. (2014: 142), on the other hand found, no relationship between male writing and prepositions. In contrast to the four predominantly male prepositions, *for* is used significantly more frequently by females in EEC80, and this aligns with Koppel et al. (2002: 408) who returned the same result with *for*. Additionally, in EEC80 both comparative and superlative adjectives are more frequently used by males. On the one hand, this is consistent with Thompson and Murachver (2001: 198) who found that adjectives generally were predominantly male features. On the other it is inconsistent with Heylighen and Dewaele (2002: 303) who found a greater proportion of adjectives of all kinds in female texts. Although the EEC80 results correspond with Thompson and Murachver (2001) with regard to adjectives, they differ in that Thompson and Murachver (2001: 198) report females asking more questions than males in their data. Men in the EEC80 use more

wh- adverbs, including *where* and *why* as well as the *wh*- pronoun *what*, which may indicate that they ask more questions than female employees (or at least ask differently). Finally, one finding that is almost unanimously agreed upon in gender profiling research is that females use pronouns more than males (Heylighen and Dewaele 2002: 303; Koppel et al. 2002: 408; Argamon et al. 2003: 325; Herring and Paolillo 2006: 449; Bamman et al. 2014: 142). Argamon et al. (2003: 332) discuss this in relation to Biber's (1988) 'involvedness' dimension, suggesting that pronouns belong to a set of linguistic features which typically signal 'interaction between speaker/writer and the listener/reader'. In contrast to these studies, pronouns return very few significant differences between male and female writers in EEC80, and those that do (reflexive pronouns and *I*) are more commonly used by men.

The fact that so many studies correlating linguistic variables with gender, including this one, produce such disparate results as to which features are characteristically 'male' and 'female', throws into question the feasibility of profiling authors by gender entirely. If the features identified were truly gender-related, then the same (or very similar), results would emerge consistently as being typical of the same group. At the same time, it is not surprising that these studies have returned different results. For one, they all examine different genres or text types: spoken data (Heylighen and Dewaele 2002), fiction and non-fiction *BNC* texts (Koppel et al. 2002; Argamon et al. 2003), emails (Thompson and Murachver 2001), blogs (Herring and Paolillo 2006), and tweets (Burger et al. 2011; Bamman et al. 2014). At the very least, these disparate findings across studies highlight the difficulty in proposing any statistical relationships between linguistic features and the gender of the author, without taking into consideration other social or contextual variables. Some of these studies have accounted for the interaction between gender and genre (e.g. Argamon et al. 2003), but others leave this entirely unexplored. Therefore, the main problem in identifying apparently causal links between gender (or any other social category) and linguistic features is that such links are often founded on: (i) relative disregard for the communicative context in which language is being produced, and (ii) over-simplifications of sex/gender and a one-dimensional view of author identity. Point (i) brings into sharp focus Johnson's (1997: 11) argument that linguists searching for differences between genders overlook the fact that all speakers and writers are drawing on the same linguistic resources. This concept of 'resources' is also discussed by Johnstone (1996: 11) in criticism of the

connections often drawn between linguistic features and social features, arguing that:

> it is more enlightening to think of factors such as gender, ethnicity, and audience as resources that speakers use to create unique voices, than determinants of how they will talk.

The influence that communicative context—whether that is the level of formality, the genre, the function or the audience—has on linguistic variation is one of the most well-established tenets of sociolinguistics (e.g. Labov 1966; Hymes 1974). More recently, in the conceptualisation of 'indexicality', Eckert (2008: 454) argues that the meanings of linguistic features are 'not precise or fixed' to particular social traits, but rather constitute a constellation of potential meanings, or an 'indexical field', 'any one of which can be activated in the situated use of the variable'. In other words, particular linguistic variables and their use are not products of, nor denotations of 'male-ness', or 'female-ness'. Instead, their meanings are determined and exploited in response to the situation in which they are used. In their study of language, gender and genre, Argamon et al. (2003: 324) acknowledge this, commenting that the differences between female and male language use 'appear to be centered about the interaction between the linguistic actor and his or her linguistic context'. These arguments underline, therefore, an inherent shortcoming in emphasising correlations between language use and gender, as findings will only ever apply to the particular context being studied. This has serious implications for authorship profiling for gender, which aims to identify overall patterns of male and female language use that analysts can rely upon when faced with a disputed forensic document.

This problem with generalising results of linguistic behaviours of males and females is related to point (ii), an oversimplification and one-dimensional understanding of author identity in profiling studies. Specifically, authors do not 'belong' to social categories. Rather, they project particular aspects of their identities in particular situations. As noted above, gender is a social construction (Eckert 1989: 253). Over the twenty-five years since Eckert's (1989) paper which first brought this debate to the fore in language research, sociolinguists have carefully examined the relationship between language and gender, and now acknowledge that 'language and discourse are meaning systems that produce (rather than reflect) gender (Weatherall 2002: 85). In other words, people use language in

such a way that *constructs* their identity (of which gender is one element), as opposed to using language in such a way *because* they are male or female. As Eckert and McConnell-Ginet (1992: 470) argue 'this implies that gender is not a matter of two homogenous social categories, one associated with being female and the other being male'. However, stylometric author profiling research does exactly this: divides people into rigid categories. As Bamman et al. (2014: 135) point out, 'there is often an implicit assumption [in author profiling] that linguistic choices are associated with immutable and essential categories of people', and this 'gives an oversimplified and misleading picture of how language conveys personal identity'. In particular, Bamman et al. (2014: 136) refer to the sociolinguistic argument that gender is socially constructed, and that it 'can be enacted through a diversity of styles and stances'. They argue that the indexical field (in Eckert's terms) of a particular linguistic feature is used by a speaker or writer to create various personae or identities, which relate to both 'global' categories like gender or race, but also to more 'local' contextual distinctions. Ultimately, they claim that 'gender and other social categories are performances, and these categories are performed differently in different situations' (Bamman et al. 2014: 138). Therefore, the finding here that there are 35 features (a combination of individual words and POS categories) which distinguish data from male employees from that of females in the EEC80 can only go so far in terms of author profiling. Sex is just one element of these authors' social identities, an element that will interact with other elements in determining the linguistic choices appropriate or required for the given situation and context. In order to adapt in such a way that is sensitive to various aspects of authors' identities, at the very least nomothetic approaches need to identify those features which relate to more than one social trait. The next section of this analysis shifts focus to the element of their identity that the authors are most likely to be performing in the unique context of EEC80: their professional identities.

7.1.2 Occupation differences

The results of the Kruskal-Wallis tests (4.4.2) find that 79 (27.15%) of the 291 features included in the analysis differentiate between occupational groups: three function word POS categories, three content word POS categories, and 73 individual function words. The Kruskal-Wallis value alone with a significance of $p<0.05$ only tells the analyst that one of the eight groups uses the feature in question with a statistically significantly higher frequency than one other group (e.g. that traders use

a feature significantly more frequently than lawyers). When such a value is returned for a feature, that feature is considered to be characteristic of the group with the highest rank mean (i.e. the group that uses it the most). In the interests of clarity of presentation, the most frequent 50% of the 79 features identified as discriminating between occupation groups are presented in Table 26.

Overall, there are significant features identified for seven of the eight occupation groups, with the only occupation group without any significant features being directors/managing directors. The group with the most features is Presidents, CEOs and COOs (=20), followed by traders (16), Vice Presidents (13), analysts, specialists and associates (11), lawyers (10), managers (5) and then finally assistants (4). There is a substantial overlap between the features which found a significant difference across genders and those which find one across occupations (marked with an asterisk in Tables 25 and 26); of the 35 features that found significant differences between the genders, 17 (48.57%) also found significant differences across occupation groups. As with gender, the vast majority of features (212 = 72.85%) show no significant difference between groups. That said, there are a greater number of features that distinguish between occupations than do so across gender, suggesting that it is easier to differentiate EEC80 authors on the basis of their occupation than whether they are male or female.

There are a number of general patterns that emerge from these results, with certain features being significantly frequent in the data of certain job groups, which could be indicative or suggestive of the differing demands, requirements and routines of their jobs, or their position in the overall hierarchy of Enron. First, different types of pronouns are dispersed across the groups. *I* and *him* are significant features for Presidents, CEOs and COOs, the group highest up the organisational hierarchy. Meanwhile, Vice Presidents prefer *we,* using it significantly more frequently, compared with *them* and *she* which are significant for lawyers and traders respectively. Traders also use indefinite pronouns generally most frequently, and in particular *anything*. A high frequency of *I* and *we* in the emails of Presidents, CEOs, COOs and Vice Presidents may indicate a more personal or involved style (Biber 1988: 105), while third person *them* and *she* in traders' and lawyers' data indicates a less personal style, making frequent reference to people outside of the immediate communicative situation. For example, compare:

Table 26. Thirty-eight features for which there is a statistically significant difference between occupation groups (Kruskal-Wallis, $p<0.05$)

Occupation	**feature**	**Sig. (p=)**	**Group mean rank**
Presidents, CEOs, COOs	total *wh-* det	.005	52.6
	total co-ord. conj	.046	57.33
	total verb (base)	.043	55.45
	an	.003	56.2
	him	.041	53.75
	*I**	.010	52.2
	*and**	.039	59.8
	*more**	.007	53.4
	your	.016	57.45
	*what**	.001	50.15
	how	.021	53.6
	*where**	.007	52.15
Vice Presidents	*we*	.015	52.5
	than	.023	54.8
	after	.025	51.2
	*my**	.004	53.5
Lawyers	*them*	.026	48.95
	of	.004	58.8
	under	.004	61.7
	no	.008	59.4
	their	.011	51.85
	which	.031	59.55
Managers	*some**	.010	49.15
	when	.020	58.9
Traders	*she*	.005	63.75
	total indef pron.	.004	58.15
	total *wh-* adverbs*	.000	56.9
	*why**	.013	54.95
	anything	.003	58.35
	because	.001	58.15
	but	.003	55.75
	might	.007	55.45
Analysts, Specialists, Associates	total comp. adj.*	.035	57.3
	total adverbs	.004	56
	into	.003	51.6
Assistants	*at*	.031	51.5
	please	.002	62.2
	thank(s)	.045	56.85

**also distinguish between genders*

(N.B. Male and female mean rank combined does not equal 100)

Example 71 (President, CEO, COO)

<From: sally.beck@enron.com>
<To: marc.eichmann@enron.com>
<Subject: Re: Commodity Logic Projects>

Thanks, Marc. **I** would like to get together with you, Mary Solmonson and James Scribner to begin work on the model for scoping high, low and expected cases around the commercialization opportunities for mid and back office.

Example 72 (Vice President)

<From: mark.taylor@enron.com>
<To: tana.jones@enron.com>
<Subject: Re: A Reminder>

We should get together with the right tax person/people - maybe Jeff can figure this out.

Example 73 (lawyer)

<From: debra.perlingiere@enron.com>
<To: karen.lambert@enron.com>
<Subject:>
Please see below. Can you help **them** with their on-line questions?
Thx
Debra Perlingiere

Example 74 (trader)

<From: chris.germany@enron.com>
<To: ed.mcmichael@enron.com, ruth.concannon@enron.com>
<Subject: RE: Assistant to print contracts>

She is going to print all the Appalachian Producer contracts that **she** can pull up out of live-link from my master list. I also asked Melissa if **she** could help and **she** said **she** would be glad too. Let me know if that's ok.

Similar differences emerge in the use of determiners. With possessive determiners, Vice Presidents prefer first person *my*, Presidents, CEOs and COOs prefer second person *your*, while third person *their* is significant for lawyers (Examples 75–77). There are also differences in quantifying determiners, with *some* being significant for managers, and *more* being significant for Presidents, CEOs and COOs (Examples 78–79).

Example 75 (Vice President)

<From: sara.shackleton@enron.com>
<To: kaye.ellis@enron.com>
<Subject: Conference Call>

Please put on **my** calendar with respect to 'POLAND'

Example 76 **(**President, CEO, COO)

<From: david.delainey@enron.com>
<To: beth.perlman@enron.com>
<Subject: ENA Offsite 2000>
<Cc: dorie.hitchcock@enron.com, tammy.shepperd@enron.com>

Beth, can you start putting together **your** presentation for the offsite. You have approximately one hour […]

Example 77 (lawyer)

<From: debra.perlingiere@enron.com>
<To: stacey.richardson@enron.com>
<Subject: Re: New GISB>

Alas it is not, I have their signature and await ENA's. Should be able to get it to you today.

Example 78 (manager)

<From: phillip.love@enron.com>
<To: greg.whiting@enron.com>
<Subject: Re: Cilco>

Thanks anyway. I have **some** information that Darron kept on Cilco […]

Example 79 **(**President, CEO, COO)

<From: mike.mcconnell@enron.com>
<To: mark.tawney@enron.com>
<Subject: 2001 goals>

Mark,
Listed below are the goals i initially submitted for your group for 2001. Please take a moment and review, change and add details to make it more clear. These worked for step 1 but i now need **more** detail - e.g. describe your long term transactions again. Please amend on this email below and return asap. […]

Conjunctions also play a role in distinguishing between different occupational groups in the sample. Presidents, CEOs and COOs use coordinating conjunctions more frequently than any other group with *and* being particularly significant. In contrast, *but* is significant for traders, as well as subordinating *because*. There is also an interesting dispersion of *wh-* words across the occupation groups. In general terms traders use *wh-* adverbs more than any other group. By looking at individual

wh- adverbs, though, we see that while traders focus on *why*, Presidents, CEOs and COOs focus on the *how* and *where* and managers focus on *when*:

Example 80 (trader)

<From: susan.pereira@enron.com>
<To: jeanne.wukasch@enron.com>
<Subject: RE: TGT cashout language>

Why are we at the 110% and not something less? When is the 30 day period over?
Susan

Example 81 (President, CEO, COO)

<From: lavorato@enron.com>
<To: david.oxley@enron.com>
<Subject:>

Where are we at with the contracts for this month.

Example 82 (President, CEO, COO)

<From: sally.beck@enron.com>
<To: hector.mcloughlin@enron.com>
<Subject: Re: Group Split>

You and Frank can best decide **how** to organize this. I am comfortable with your decisions, so feel free to go ahead with this. --Sally

Example 83 (manager)

<From: mike.grigsby@enron.com>
<To: dale.neuner@enron.com, melba.lozano@enron.com>
<Subject: Baja and redwood spreads>

Please contact Keith Holst if you should have any questions regarding the new spread products. Please let him know **when** they are completed. I will be out of the office until Wednesday. Thank you very much.
Mike Grigsby

Total *wh*- pronouns and determiners are also significant for Presidents, CEOs and COOs, with *what* being particularly important, compared with *which* as a preference of lawyers. There are only four features in total that were significant for assistants, the employees lowest in the hierarchy. Two of those are politeness features of *please* and *thank*, which could be explainable by the fact that whenever they are emailing, they are corresponding up the organisational hierarchy:

Example 84 (assistant)

<From: sherri.sera@enron.com>
<To: greg.piper@enron.com>
<Subject: Re: FW: eWorldTradeX>
<Cc: philippe.bibi@enron.com, tina.spiller@enron.com>

Greg, I had a message to call Joe Dial yesterday. I haven't called him back yet. Should I wait to hear from you? **Please** advise. **Thanks**, SRS

Finally, analysts, specialists and associates use both comparative adjectives and adverbs statistically significantly most frequently:

Example 85 (analysts, specialists and associate)

<From: susan.scott@enron.com>
<To: glen.hass@enron.com>
<Subject: Negotiated Rate>
<Cc: mary.darveaux@enron.com>

Glen, as we discussed, here is the negotiated rate TW deal which TK did today under Rate Schedule ITS-1 [...] We probably need to put our heads together to come up with a **better** procedure for when this happens. It would be so much **better** if we were able to send tariff sheets via e-mail to our D.C. office. Call if you have questions or need anything.

Example 86

<From: susan.scott@enron.com>
<To: tony.pryor@enron.com>
<Subject: thoughts on Enrononline>

Tony, [...] **Probably** the best we can do is extrapolate from 1) the tariff and 2) the NGA (doesn't it also say we have to provide service on a not **unduly** discrim. basis [...]

The fact that so many features statistically discriminate between both gender and occupation highlights the difficulty in divorcing one aspect of an author's identity from any other. For instance, the first person personal pronoun *I*, is used significantly more frequently by males than females, but it was also significant for Presidents, CEOs and COOs. However, eight of the ten people constituting this group were males, and this complicates the results; it is difficult to know whether the feature is a true discriminator of gender, or a true discriminator of occupation, or neither. In the same way, *wh*- adverbs generally are a male feature in the first instance, but as the results above show, they are also significant for traders, while specific *wh*-adverbs (*why, when, where, how*) were significant for different groups altogether. This was the case for a considerable number of features in this analysis.

Therefore, we cannot be sure whether these words are products of individuals performing gender roles or occupational roles. A possible solution to this would have been to have a male and female category for each occupation type. This way, 'male traders', for example, could have been compared with 'female traders'. This approach is very difficult to apply in the profiling of Enron authors, however, given that one of the gender is very sparsely represented in a number of occupation groups, such as the heavily male-dominated President, CEO and COO group, or the female-exclusive assistant role (see Section 3.4.4). Nevertheless, these overlaps of results between gender and occupation provide evidence to suggest that correlations between linguistic features and any one social characteristic are not reliable, and should not be considered definitive or conclusive, without taking into consideration other potential influences on the use of the features in question. Indeed, in a sociolinguistic context Baxter (2012: 102) notes the complex relationship between language, gender and occupational role in an organisational context, finding that individuals adopt particular linguistic strategies to accomplish communicative goals in such a way that are difficult to account for in terms of either occupational role or gender alone. Therefore, as has been argued throughout this thesis, relying on quantitative results alone does not suffice in the analysis of linguistic style, either of groups or individuals. Nomothetic profiling results as presented here, and throughout the profiling literature, provide the analyst with only superficial evidence about the importance of particular words in identifying stylistic differences between groups of authors. The analysis that follows focuses on three features that have been identified above as discriminating between gender, occupations or both in EEC80: *for*, a significantly female word, *hi*, also a significantly female word, and *why* a significant word for both males and traders. The analysis shifts from the overall nomothetic results presented here to considering the use of these words in their linguistic and communicative contexts by the different groups. Furthermore, adopting an idiographic approach through the focus on specific words allows for a comparison across all of the individual authors making up the gender and occupation groups. Such an approach can be used to examine how individual authors exhibit distinctive preferences in using the features in question, a phenomenon that is overlooked by relying on overall statistical group-level findings. In particular, this analysis shows the ways in which an idiographic approach can supplement the nomothetic results, and help us better understand the relationship between people's

linguistic choices and the various elements of the multi-dimensional identities they perform and project.

7.2 Is *for* a 'female' feature?

In their author profiling research in which they aim to identify the gender of writers, Argamon et al. (2003: 323) comment that their 'main interest is to present the linguistic phenomena' and that they endeavour 'to avoid baseless speculation with regard to interpretation of the data'. However, Grant (2008: 226) warns against computationally pursuing an algorithm which distinguishes between authors, but has no linguistic explanation or validity, and if the aim of author profiling is to reliably aid forensic investigation (Grant 2008: 224) then the presentation of statistical findings is not enough. Further linguistic analysis is necessary to supplement and explain statistical results. The case of *for* is used here to demonstrate that linguistic exploration and interpretation of a feature which is found to distinguish between groups need not be considered 'baseless speculation'. Rather, it can reveal that the quantitative differences observed are indicative of underlying differences in the linguistic strategies and preferences of different groups of authors.

For has been selected for this analysis given that in the Mann-Whitney U tests reported above it was found to be used by females statistically significantly more than males, and no significant difference was found across occupations. The Mann-Whitney U test is based on ranks rather than frequencies themselves (see Section 4.4.2). Therefore, while the test identified a statistically significant difference between males and females in terms of ranks, the two groups actually use *for* with very similar relative frequencies; 12,304 instances of the preposition account for 1.23% of all tokens in the female data, while 14,154 instances in the male data account for 1.20% of all tokens. However, the differences between male and female authors are greater when we consider that seven of the top ten users of *for* are female. Despite the difference in frequency of use, the L1 and R1 collocates which the two genders co-select with *for* are very similar (Table 27).

Table 27. Top ten L1 and R1 collocates of *for* in male and female authors in EEC80 (% of all instances of *for*).

L1 collocates					
males (*for* =14,154)			**females (*for* =12,304)**		
thanks	1,051	(7.43%)	*thanks*	950	(7.72%)
you	205	(1.45%)	*you*	193	(1.57%)
looking	179	(1.26%)	*looking*	142	(1.15%)
up	175	(1.24%)	*up*	133	(1.08%)
sorry	154	(1.09%)	*agreement*	129	(1.05%)
works	123	(0.87%)	*work*	113	(0.92%)
out	117	(0.83%)	*list*	111	(0.90%)
work	106	(0.75%)	*this*	102	(0.83%)
much	101	(0.71%)	*works*	102	(0.83%)
pay	85	(0.60%)	*sorry*	101	(0.82%)
R1 collocates					
males			**females**		
the	3,148	(22.24%)	*the*	2,034	(16.53%)
a	660	(4.66%)	*your*	787	(6.40%)
your	645	(4.56%)	*a*	564	(4.58%)
me	470	(3.32%)	*me*	522	(4.24%)
you	385	(2.72%)	*you*	462	(3.75%)
this	331	(2.34%)	*this*	316	(2.57%)
all	205	(1.45%)	*all*	263	(2.14%)
us	157	(1.11%)	*us*	185	(1.50%)
my	118	(0.83%)	*financial*	106	(0.86%)
an	114	(0.81%)	*my*	104	(0.85%)

Although there are some differences in the lists, such as *much* and *pay* being top L1 collocates for males and *agreement* for females, it is the similarity between the two which is most striking. For instance, the top four most frequent L1 collocates—*thanks, you, looking* and *up*—are the same for both men and women. The similarities are even more consistent in R1 collocates, nine of the top ten collocates for both groups are the same, and almost in the same order of frequency (*a*, *your* and *my* being minor exceptions).

This similarity across collocates is indicative of very frequent word n-gram clusters occurring in the corpus, used by both men and women, such as:

- *thanks for the* (n=793)
- *thanks for your* (n=663)
- *thank you for* (n=252)
- *sorry for the* (n=214)
- *for you to* (n=187)
- *for me to* (n=156)

These n-grams might be considered as being typical or conventional in the (business) email genre. For example, *thanks* is by far the most frequent L1 collocate of *for* in the EEC80, accounting for 2,001 (7.56%), but only 12,130 (0.31%) of the 3,934,071 *for* occurrences in *COCA*. Similarly, *sorry* is the fifth most common *for* L1 collocate in EEC80 (255 = 0.96%) but is far less frequent in *COCA* (4,078 = 0.1%). Overall, then, the use of *for* is similar for males and females, and is found in commonly shared, generic patterns which appear typical of this kind of communication. That said, the *way* in which these word n-grams are employed by males and females in the EEC80 is different. In fact, more optimistically, they may provide useful evidence for determining not only the gender of Enron authors, but also their occupation.

Differences can be drawn between male and female authors here on the basis of how *for* is used within the most common collocation: *thanks for*. Of the 2,001 instances of *thanks for* in EEC80, 793 (39.63%) are followed by *the* and 663 (33.13%) are followed by are followed by *your*. Combined, the trigrams *thanks for the* and *thanks for your* account for 72.76% of all instances of *for* in EEC80. Before discussing the exact nature of the differences between males and females in their use of *thanks for*, it is worth noting the different communicative functions of *thanks for the* and *thanks for your*. The former is generally used by EEC80 authors to thank their recipients for some prior communication between the two, either an *update* (n=129), *info(rmation)* (91), *email* (52), *invite/invitation* (56), *note* (44), *heads up* (39), *message* (25), *offer* (24), *feedback* (16), *reminder* (16) (Figure 27).

Figure 27. 25 of 793 concordance lines for *thanks for the* in EEC80

N	Concordance
1	(Fax) <Message-ID: > Dan: Thanks for the e-mail. Things are going
2	looks great. I hope all is going well and thanks for the e-mail. Talk to you soon.
3	Yes, thanks. m <Message-ID: > Cindy, Thanks for the email. I 'll be glad to set it
4	> Congrats! Mark <Message-ID: > Don, thanks for the feedback. I am supposed
5	Regards, Mark <Message-ID: > John, thanks for the feedback on Mark and
6	thoughts...Thanks! <Message-ID: > Thanks for the heads up! As usual,
7	is mandatory. Mark <Message-ID: > Thanks for the heads up. Keep me
8	going until 330. Kay <Message-ID: > Thanks for the info. I'll give it a look. Kay
9	have any questions. <Message-ID: > Thanks for the info. <Message-ID: >
10	ask about that. EB <Message-ID: > Thanks for the information. We are using
11	Vince <Message-ID: > Michael, Thanks for the information. We shall
12	FYI Vince <Message-ID: > Vasant, Thanks for the invitation. It works for me.
13	> OK by me. <Message-ID: > Thanks for the invite (and the follow up
14	fix this for me. Mark <Message-ID: > Thanks for the message. I will try to get
15	about it. Vince <Message-ID: > Dave, Thanks for the message. I don't think we
16	Mark <Message-ID: > Susan: Thanks for the note. Glad to hear that
17	Thanks, Kim. <Message-ID: > Earl, Thanks for the note. Mansoor and I have
18	Right! Thanks! <Message-ID: > No, but thanks for the offer. Ben <Message-ID: >
19	> dont know them <Message-ID: > thanks for the offer. i'll let you know
20	Roger the docs separately, but thanks for the reminder. I am capable of
21	desk and we'll sort it out this afternoon. Thanks for the reminder. Kate
22	I will bring handouts. <Message-ID: > thanks for the update. This is very useful.
23	> Leslie, This looks good to me. Thanks for the update. Bill <Message-ID:
24	by the gas group will need to be moved. Thanks for the heads-up. Stacey
25	Thanks. Mark <Message-ID: > Martin: Thanks for the update! It looks like great

Such reference to other communication highlights the important role of intertextuality in workplace discourse. Koester (2010b: 41), for example, notes that intertextuality in the workplace is primarily manifest in 'many explicit references in the discourse used by the participants to other discourse acts'. This use of *thanks for the* by employees in EEC80 is evidence of such explicit reference to other discourse acts. In some cases, the reference constitutes the whole email message (Examples 87–88), while in others the author elaborates either briefly (Example 89) or extensively (Example 90).

Example 87

<From: john.lavorato@enron.com>
<To: michael.guerriero@enron.com>
<Subject: Re: Argentine Transaction Summary>

Mike
Thanks for the info.

Example 88

<From: phillip.allen@enron.com>
<To: wise.counsel@lpl.com>
<Subject: RE: Huntley update>

Thanks for the update.

Example 89

<From: jeff.skilling@enron.com>
<To: nasim.khan@enron.com>
<Subject: Re: Online Foreign Exchange>

Thanks for the e-mail, Nasim. We're working on a variation of it now. Keep up the good work.
Jeff

Example 90

<From: jim.steffes@enron.com>
<To: tom.hoatson@enron.com, sarah.novosel@enron.com, daniel.allegretti@enron.com,>
<howard.fromer@enron.com, l.nicolay@enron.com>
<Subject: RE: Enron ICAP Strawman>
<Cc: steve.montovano@enron.com>

Thanks for the feedback.

Your model now assumes a modification to the Distribution tariff to include a new rate element. That may get around the Settlement (although OCA in PA will think that now D is going up and there is no offset in the G rate so consumers are getting screwed because they won't switch)[...]

Thanks for your is sometimes used in a similar way, being followed by *note* (n=43) and *message* (n=28), but its use in EEC80 is reserved primarily for thanking recipients for their *help* (372) and *assistance* (24) (Figure 28). The frequency of either the sender's name as a sign-off or <Message-ID:> indicating the start of a new message directly following *thanks for your help/assistance* shows that these sequences often occur at the end of emails, in anticipation of compliance on the part of the recipient. Indeed, 70.45% (279/396) of the occurrences of these phrases appear at the end of the email, either in the place of, or directly before, the sign-off:

Figure 28. 20 of 663 concordance lines for *thanks for your* in EEC80

N	Concordance
1	this will make for a smooth process! Thanks for your assistance.
2	contact and I will start over with EOL. Thanks for your assistance in this
3	as changes to these agreements. Thanks for your assistance, Kay
4	agreement and the GE parent guaranty. Thanks for your assistance. The
5	Thanks, Matt Lenhart <Message-ID: > thanks for your help. <Message-ID: >
6	days and data for imports/exports. thanks for your help. matt <Message-ID:
7	is the division of transcanada. thanks for your help. matt <Message-ID:
8	a resignation letter from C. Supatgiat. Thanks for your help today. Vince
9	pulp & paper description for our group. Thanks for your help! <Message-ID: >
10	Or do you want me continue to say no? Thanks for your help! <Message-ID: >
11	be of any benefit? <Message-ID: > Thanks for your help! Are you getting
12	Bruce Mills - bmills Phillip Love - plove Thanks for your help. PL <Message-ID:
13	now and can't take care of this myself. Thanks for your help! Mark T.
14	are on a conference call now re funding. Thanks for your help, Kay <Message-ID:
15	of my frustration with them out on you. Thanks for your help and thanks for not
16	probably also needs the final itineraries. Thanks for your help. Mark T.
17	Right??????? <Message-ID: > Not yet. Thanks for your help. <Message-ID: >
18	we can get these contracts assigned. Thanks for your help. Daren
19	meeting with them on Wednesday. Thanks for your help, Kim <Message-ID:
20	and getting them to Bear Stearns. Thanks for your help! Marie

Example 91 (trader)

<From: eric.bass@enron.com>
<To: gary.taylor@enron.com>
<Subject: JAN-MAR Degree Day Swaps>

Gary,
What is the market for IAH Jan-Mar HDD swaps (we are looking to BUY)? How does this compare to the 30 yr avg?
Thanks for your help.
Eric

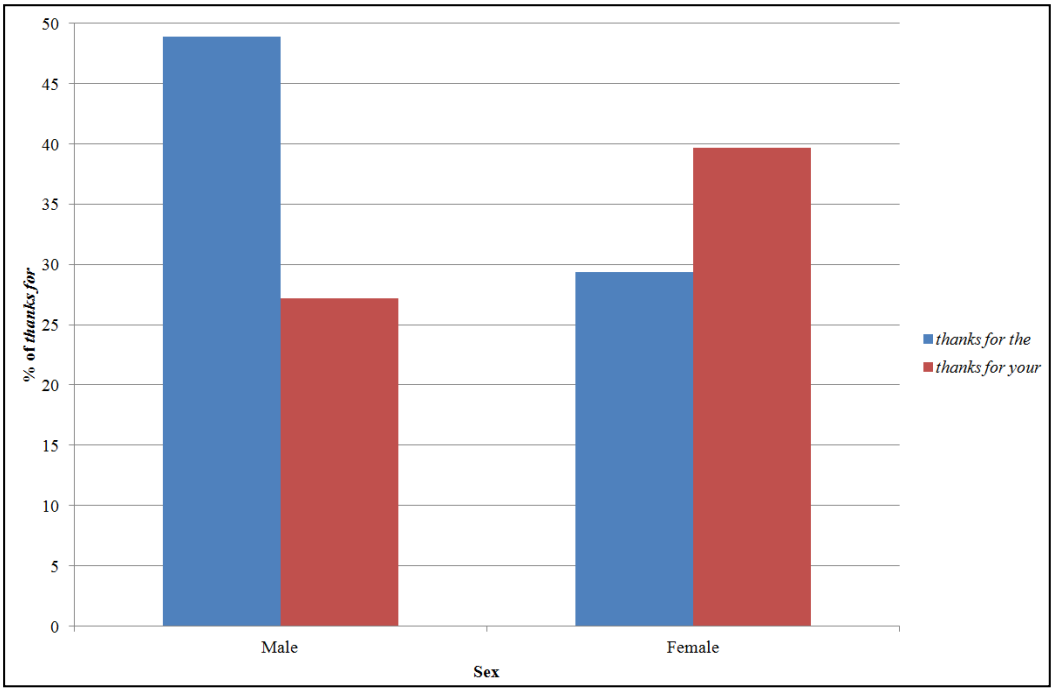
Example 92 (lawyer)

<From: gerald.nemec@enron.com>
<To: mary.ogden@enron.com>
<Subject: RE: Energen Resources CA/Lewis Energy CA>

Mary, Could you please prepare the form of CA for these entities. Please make them bilateral. Also check with Kay Young for conflicts. Let me know if you have any questions. **Thanks for your** help.

The major distinction that can be drawn between male and female EEC80 authors in terms of *for*, is through the frequencies with which they use these *thanks for the* and *thanks for your* n-grams (Figure 29). Males use *thanks for* a total of 1,051 times, 514 (48.91%) of which are part of *thanks for the* compared with only 286 (27.12%) *thanks for your*. The pattern is the reverse for women; *thanks for* occurs 951 times in the their data, 377 (39.64%) of which are *thanks for your*, compared with only 279 (29.34%) *thanks for the.*

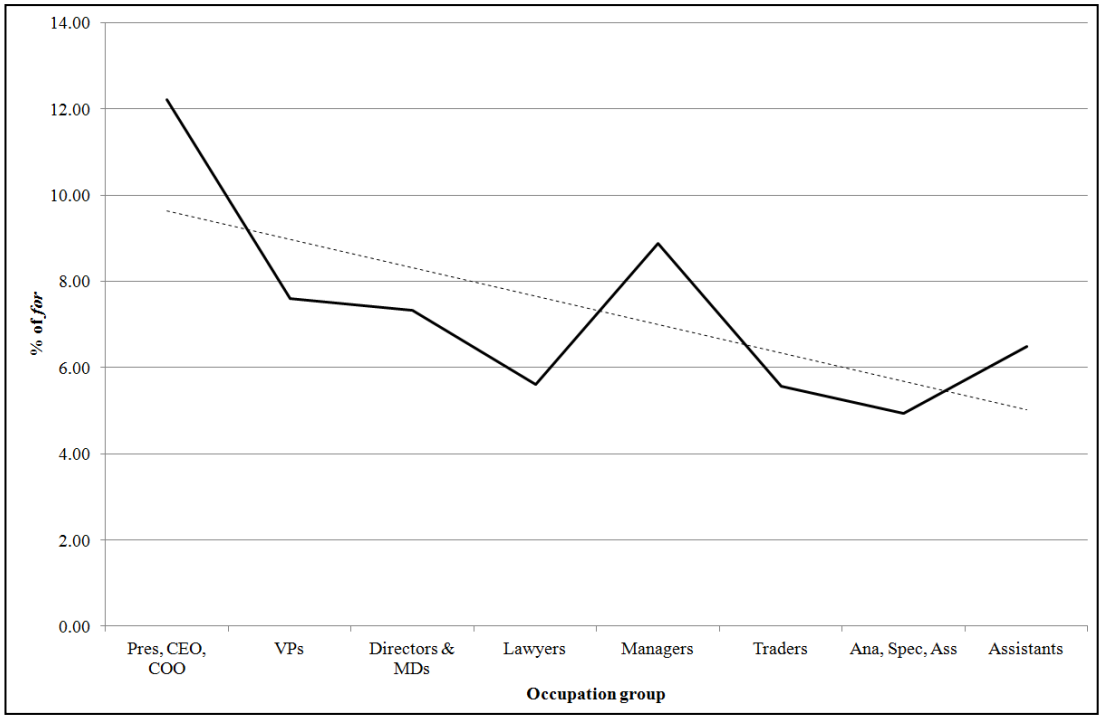
Figure 29. Comparing *thanks for the* and *thanks for your* by gender in EEC80



This finding can supplement the nomothetic result that females use *for* more than males. Analysing its use in its most common collocation patterns has identified a major underlying stylistic difference between men and women. Male uses of *for* are more frequently part of explicit intertextual references to other discourse acts and thanking recipients for previous communications. In contrast, females more frequently use *for* in thanking recipients for their *help* and *assistance*, most often at the end of an email. Such behavioural patterns offer richer more sociolinguistically explainable, yet still quantifiable, differences between the genders than general nomothetic results do. It might be, then, that the profiling of gender using collocational analysis such as this holds more potential than the comparison of individual words.

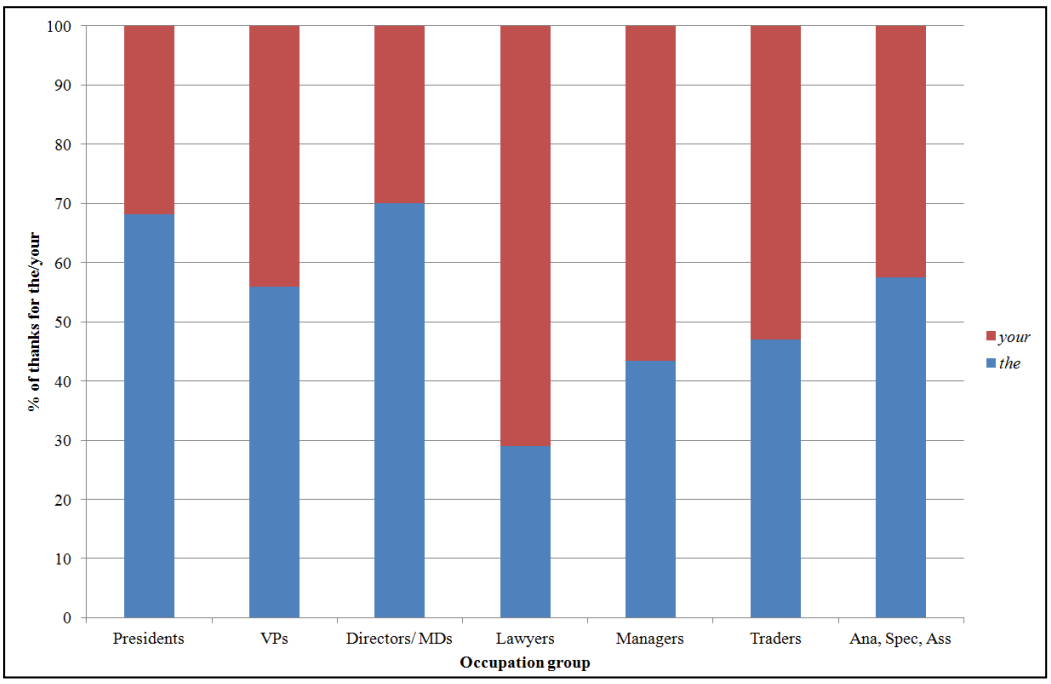
An analysis of these n-grams across occupation types, however, suggests that the primary influence on the use of these n-grams is occupation, not gender. Returning to the use of the *thanks for* bigram, differences emerge across occupations. In particular, the general trend is that those higher in the Enron hierarchy use *thanks for* more frequently than those lower down the hierarchy, ranging from 12.21% of all *for* instances for Presidents, CEOs and COOs, to 4.93% for analysts, specialists and associates (Figure 30).

Figure 30. *Thanks for* frequencies across the EEC80. *(--- trend line)*



This indicates that those with more managerial responsibility, such as Presidents, CEOs, COOs, Vice Presidents, directors and managing directors may be receiving more updates, emails, messages, feedback and help for which they need to thank their recipient. The main exception to this general trend is that middle managers are the second most frequent *thanks for* users, exceeded only by Presidents, CEOs and COOs. In turn, the use of *thanks for* may be considered sociolinguistically constrained (Turell 2010: 239) by individuals' professional identities, and the roles they are fulfilling in the community of practice, rather than their gender. Furthermore, differences can be drawn between groups, as the frequency of *thanks for the* and *thanks for your* are compared (Figure 31).

Figure 31. *Thanks for the/thanks for your* compared across occupations in EEC80. (Assistants excluded as *thanks for* frequency only 28)



While higher ranking employees prefer *thanks for the*, relating to previous communication, *thanks for your*, relating to the acknowledgment or expectation of help, is preferred by lower ranked groups. These patterns across jobs go some way towards explaining the differences across the gender. *Thanks for the* is used more by males and those in the top three ranking occupation groups—Presidents, CEOs and COOs, Vice Presidents and directors/managing directors—as well as analysts, specialists and associates who are also frequent in their use of this n-gram. Of the 40 authors who make up these four groups, 27 are male. In contrast, females prefer *thanks for your* and the results in Figure 31 show that it is more common in lawyers than any other occupation group. This could be related to lawyers drafting, reviewing and redrafting legal documents, and often thanking colleagues for their help in this practice (see also the case study of Gerald Nemec in Section 6.4 above). However, *thanks for your* is used a total of 164 times by only five of the ten lawyers, four of whom are female: Tana Jones (72), Debra Perlingiere (33), Marie Heard (17) and Kay Mann (11). The only male lawyer to use this n-gram is Gerald Nemec (31). In these cases it is difficult to ascertain whether the use of this feature is mostly determined by the authors' gender or their occupational identity, or even whether it is a personal stylistic choice and a habitual element of their idiolects. For example, is *thanks for the* identified as a feature of the more powerful ranking employees

because it is a feature associated with their roles in the Enron community of practice, or is it truly a 'male' feature, appearing to be related to these occupation types simply because most of the authors in these positions are male? The same is the case with *thanks for your*. Is this a lawyer feature, or does it simply appear to be, given that the most frequent users in this group are all women? This inability to determine which aspect of the author's identity has most influence on linguistic production is only a problem from the perspective of nomothetic profiling that aims to quantitatively correlate linguistic features with individual social categories, divorcing the different aspects of a person's identity. For many years sociolinguistic research into gender and occupational status has readily identified the ways in which these different social variables influence linguistic practices in the workplace (e.g. Woods 1988; Kendall and Tannen 1997; Eckert and McConnell-Ginet 2003; Holmes 2006; Talbot 2010) and how they interact in 'claiming, negotiating and renegotiating our emerging identities in interaction' (Angouri and Marra 2011: 1). Therefore, a conclusion here might be that while *thanks for the* is an n-gram particularly favoured by male employees within higher ranking occupational roles in Enron, *thanks for your* is most common in the writing of female lawyers.

The importance of this kind of conclusion transcends this particular study of Enron, as it highlights the value of taking a more multi-dimensional approach to identity than is currently practiced in author profiling research. The original statistical tests at the start of this chapter identified *for* as being significant for females, but found no such significance for any occupation group. A subsequent stylistic analysis of how *for* is used within its most common n-grams across the EEC80 authors has identified an important relationship between *for*, gender and occupation. This relationship was overlooked by the initial statistical results. Therefore, profiling of authors in this way perhaps offers more promise than the purely quantitative approach. The danger of quantitative author profiling in its current form is that apparent correlations revealed between certain linguistic features and social categories may be caused by another aspect of either identity or communicative context that has been overlooked because the difference between groups was not 'statistically significant'. At the same time, this analysis demonstrates the value of examining further the linguistic features identified as being significant in distinguishing between groups of people. The finding that women use *for* more than men is limited in its usefulness. However, the grammatical

nature of *for,* and of other function words identified as useful in author profiling, is such that it is indicative of underlying lexico-grammatical choices made by authors. A stylistic analysis of how these words are used in collocations and co-selections across different groups of authors can help elaborate on the statistical findings, and ultimately identifies further linguistic differences between groups.

7.3 *hi* and the comparison of two female lawyers

In the same way as *for, hi* was identified by the Mann-Whitney U tests as being a word used significantly more by females than males in EEC80. It has been chosen here for a more in-depth analysis, given that it serves an important pragmatic function as an email greeting in workplace email discourse. In addition, previous research has found that email greetings used in the Enron Email Corpus can be very distinctive of individual authors (Wright 2013), but no discussion has yet been had over the different greeting preferences across groups in the corporation. Finally, greetings often represent 'smoking guns' of authorship, and variation within greetings has been used as a marker of style in attributing authorship of emails (e.g. Turell 2010; Grieve 2013). This section analyses closely the use of *hi* in greetings, but does not take into account corresponding email sign-offs, which may provide additional communicative differences across groups; Wright (2013) found distinctive variation in email openings and closings between individual Enron employees. The particular focus here is on the extent to which the difference in frequency of *hi* is a product of different underlying pragmatic strategies employed by the different gender in the negotiation and projection of their institutional identity when emailing recipients with a specific purpose.

There are a total of 789 instances of *hi* in the EEC80 sample, 612 (77.57%) of which are used by female employees, accounting for 0.06% of the 996,796 words in the female sub-corpus. This is compared with 177 occurrences in male data, accounting for only 0.01% of their 1,181,409 tokens. Only 46 of the 80 authors in the sample use *hi,* of which 25 are female (from a total of 34 female authors in the sample) and 21 male (from a total of 46). There are a total of 23,546 emails sent by female employees in the EEC80 sample, 2.6% (612/23,546) of which are opened by the author greeting the recipient with *hi*. This makes *hi* an intriguing feature as even though it is a relatively infrequent choice in the sample, females still use it

significantly more frequently than males. In both the female and male data, *hi* is most frequently followed by the recipient's first name (Figure 32).

Figure 32. 10 of 584 concordance lines for *Hi* + name in EEC80

N	Concordance
1	FYI. We have this report? <Message-ID: > Hi Aleck: How did your talk go? I hope well.
2	print <Message-ID: > Hi Andy, I would like access to the ICEX--1
3	at on Friday. Best, Jeff <Message-ID: > Hi Barry: Congratulations on side-stepping
4	Sorry, you are on the list. <Message-ID: > Hi Betsy, Sorry for the delay. What are you
5	yet for the call? Best, Jeff <Message-ID: > Hi Bob: Could you please leave on my
6	through Blythe. Best, Jeff <Message-ID: > Hi Brad: I finally got in touch with the people
7	p 28 o 20 mon p 45 o 20 <Message-ID: > Hi Cheryl, I'm not sure if I forwarded this or
8	what he is interested? Jeff <Message-ID: > Hi Chris, Give me a ring tomorrow (tuesday)
9	it's enron's phone, so ... <Message-ID: > Hi Chris, I don't know where you're at with
10	please advise. Cooper <Message-ID: > Hi Chris, I'm deeply concerned about all of

This pattern is far more commonly used by females, with 491 (80.23%) of their 612 *hi* instances being followed by a first name, in comparison with only 93 (52.54%) of 177 occurrences in the male data. Instead, males choose to have *hi* as a standalone greeting with no name, or refer to multiple recipients using terms such as *gang*, *guys* or *team* (Figure 33).

Figure 33. 10 of 177 concordance lines for *Hi* in male EEC80 data

N	Concordance
1	PCG is up $1.6 as we speak. <Message-ID: > Hi. Any word on the mower? <Message-ID: > very
2	said he has not seen them. <Message-ID: > Hi gang. You guys may or may not be interested in
3	not above groveling for affection. <Message-ID: > Hi gang. Please read the following and let me know
4	Thanks guys. Any ideas? <Message-ID: > Hi guys, I've had a cancellation for tomorrow's
5	coordinate with you. Later Jeff <Message-ID: > Hi guys, Shawn Cumberlan, the former head of the
6	of weeks ago? Thanks, Mike <Message-ID: > Hi, I noticed that there are several people missing
7	some ideas and information, too. <Message-ID: > Hi: Sorry to bug you--know things are crazy--but
8	- Market East. Thanks dude! <Message-ID: > Hi Team! I just extend CNG deal 116090 through
9	paid ENA for October activity? <Message-ID: > Hi team. Do not discuss this with anyone at CES
10	craziness out there. Best, Jeff <Message-ID: > Hi. We'll need to add this to the update. Hot off the

In research on workplace discourse, greetings have typically been considered 'phatic' in nature. When describing the generic structure of service encounters, Hasan (1985) categorises greetings as 'optional elements' and McCarthy (2000: 104) labels greetings and partings in salon and driving lesson interaction as 'phatic exchanges'. Holmes (2000: 38) suggests a continuum of task-orientation for workplace communication, in which 'core business talk' is at one end and 'phatic

communication' is at the other. In discussing Holmes' model, Koester (2006: 55) explains that 'core business talk is "on topic" in terms of the transactional goal of the particular encounter' whereas 'phatic communication is topically unrelated to the workplace'. However, Koester (2004; 2006; 2010b) highlights the difficulty in disconnecting transactional from 'relational' interaction in the workplace, and emphasises that 'it is not possible to neatly separate talk that is purely instrumental from talk that has a relational or social purpose' (Koester 2010b: 97). Research such as Bou-Franch (2006; 2011), Waldvogel (2007), Bjørge (2007), Economidou-Kogetsidis (2011: 3205) and van Den Eynden (2012) has focused on the interactional function of greetings in institutional emails. Bou-Franch (2006: 82), for example, argues that email greetings are moves 'of great social significance' through which participants 'seek common ground' and 'emphasize that both co-participants belong to the same community of practice'. Kankaanranta (2005: 359) notes that by using email greetings, a writer 'constructs a relationship with the recipient, and the usage thus contributes to the maintenance of good social relations'. Similarly, Waldvogel (2007: 457) argues that email greetings in the workplace are 'one means by which the writer constructs his or her social and professional identity and relationship with the addressee(s)'. Therefore, it can be argued that greetings in workplace emails serve a transactional purpose, rather than being relegated to 'optional' or 'phatic' elements.

Waldvogel (2007: 463) proposes a relationship between the type of greeting used and the transactional or functional role of the email. In particular, she draws attention to the *hi* + name greeting, the sort that is so prolific in the writing of the female authors in the EEC80. She argues that the use of a greeting word (e.g. *hi*) and the recipient's name is utilised in emails in order to 'introduce a matter of a fairly delicate nature' such as 'a major request'. Greeting the recipient with *hi* and their first name, such as in Figure 32, constitutes 'vocative naming' or 'a proper name used in address' (Wales 2001: 406). Vocative naming is commonly used with imperative or directive utterances (Downing 1969; Kleifgen 2001), that is, utterances which 'attempt to get someone to do something'. Indeed, in the EEC80 sample, the *hi* + name greeting is found in two different types of emails: those in which the sender is requesting information or giving a directive, and those which serve an information-giving function. Of the 491 instances of *Hi* + name in the female data, 276 (56.21%) occurred in emails in which the sender either requested

information from the recipient or gave a directive (Examples 93–94). In contrast, only 25 (26.88%) of the 93 *hi* + name instances used by males are found in such emails. Instead, males more frequently use this greeting form in information-giving emails (Example 95 and 96). Therefore, males and females use *hi* + *name* with different frequencies, and they also use it with different functions.

Example 93

<From: kay.mann@enron.com>
<To: chip.schneider@enron.com>
<Subject: Delta - Hancock letters>

Hi Chip,
Have you seen anything from Hancock re Delta? I haven't.
Thanks,
Kay

Example 94

<From: kay.mann@enron.com>
<To: carlos.sole@enron.com>
<Subject: THE agreement>

Hi Carlos,
Could you please get Dick to sign this? Can he sign for NEPCO also? Two original sigs is enough.
Thanks,
Kay

Example 95

<From: vince.kaminski@enron.com>
<To: energy.vertical@juno.com>
<Subject: RE: Jeff Skilling resigns>

Hi Joe,
I called you Tuesday and left a message on your voice mail.

Example 96

<From: jeffrey.shankman@enron.com>
<To: john.sherriff@enron.com>
<Subject:>

Hi John, hope things are well...Hickerson and I were talking yesterday about the convert bond desk, and he mentioned to me that it make sense for it to come to his group [...]

A number of studies have identified the *Hi* + name email greeting as being an informal choice. The use of first names in addressing recipients reflects social familiarity between interactants (Laver 1981: 299; Holmes 2001: 267–72) and Biber

and Conrad (2009: 189) found that this particular greeting was most commonly found in emails between friends and family. In workplace contexts, this greeting is used more commonly where there is a smaller 'power distance' between communicants (Bjøge 2007: 72), or when a writer wants to reduce the hierarchical distance between high and low ranking participants, as well as in emails with a transactional purpose. It 'creates a greater sense of solidarity' between employees and suggests that 'people of lower status are acknowledged and treated with respect' (Waldvogel 2007: 467). In the EEC80 sample, then, when female employees frequently use this greeting, they are doing so in such a way that mitigates the request or directive being made. Using an informal and friendly greeting serves to reduce the social distance between sender and recipient, and as such reduces the face-threat of the request or directive. This is a strategy they use equally when writing to males and females; 143 (51.81%) of the 276 instances in which they use *Hi + name* before a request are sent to male recipients, and 133 (48.19%) are sent to females. In some cases, the *Hi* + name greeting is used in combination with other informal or non-transactional strategies, including humour, sarcasm, self-deprecation (Example 97–100).

Example 97

<From: kay.mann@enron.com>
<To: chris.gaffney@enron.com>
<Subject: Asset management>

Hi Chris,
Happy New Year. Hope all is going well with you. We miss you!
I heard a rumor that the UI deal had an asset management aspect to it. The guys around here are kicking around some different asset management ideas, so I'm trying to get my hands on any docs which might give some insight into some different approaches. Do you have anything you can email me?
Thanks,
Kay

Example 98

<From: elizabeth.sager@enron.com>
<To: chris.gaffney@enron.com>
<Subject: hi>

Hi Chris
Hope all is well with you, Tracy and the kids. How is Addy (opps, spelling is bound to be wrong)? Are you coming to the legal conference? And now for the real reason of my email - Can you send me Tracy's telephone number. I've been wanting to call her for months and say hi but the stupor got in my mind and I couldn't shake it free. Its a beautiful day here today - the first for months and I'm committing to a new .... well, at least a call to Tracy. Hope to see you at the legal conference. Did you hear I have to speak? At least it is with Klauberg.
Elizabeth Sager

Example 99

<From: marie.heard@enron.com>
<To: diane.anderson@enron.com>
<Subject: Catequil>

Hi, Diane:
In your spare time (HA! HA!) will you check and let me know the earliest outstanding trade dates for the following counterparties:
Catequil Overseas Partners, Ltd.
Catequil Partners, L.P.
They are ready to sign the ISDAs.
Thanks!

Example 100

<From: marie.heard@enron.com>
<To: credit.williams@enron.com>>
<Subject: Southern California Gas and Petro-Hunt>

Hi, Jay:
It's your friendly pest AGAIN!!!!!!!!
Do you have an address for me for Southern California and have you found out anything about 'Petro-Hunt L.L.C.' and 'Petro-Hunt Corporation being the same entity?
Thanks!
Marie

In Example 97, Kay Mann engages in small talk—social or casual conversation (Schneider 2008: 102)—with her recipient, wishing him a *happy new year*, enquiring about his well-being and complimenting him. In Example 98, Elizabeth Sager's request for a telephone number is embedded within a number of 'relational episodes', small talk occurring during the performance of a transactional task (Koester 2004: 1420). She goes as far as to signpost the point at which the purpose of the email emerges with *And now for the real reason of my email*. In Examples 99 and 100, Marie Heard uses humour to preface directives. *Ha! Ha!* is used to

acknowledge that the recipient—Diane Anderson—is unlikely to have much *spare time*. This acknowledgement reduces the force, or at least the urgency, of the directive which immediately follows. Similarly in Example 99, Heard refers to herself as *your friendly pest*. This self-deprecation seemingly refers to the fact that she is regularly making requests or directives to this particular recipient. The use of the possessive determiner *your*, the adjective *friendly*, capitalisation and multiple exclamation marks all contribute to an informal style, before the double request which follows. Humour used in the workplace in this way builds a positive sender identity and shows solidarity and convergence with the recipient (Holmes and Marra 2002; Koester 2010b: 112). These non-transactional mitigating strategies are all used in combination with the informal, familiar *Hi* + name greeting to preface requests and directives. Given that males use *hi* + name far less than female authors in EEC80, and use it more frequently in information-giving emails than requests or directives, these mitigating strategies identified here are distinctive of women in this sample. This partially supports the results of Waldvogel (2007: 463) who found that in an educational organisation females use greeting word + first name more than males, but found the opposite pattern in a manufacturing plant, which may be due to the overarching gender-bias of the different environments. This has implications for the status of *hi* as a significant feature in the profiling of Enron employees. The fact that *hi* is used significantly more frequently by female than male authors in the EEC80 appears to be a quantitatively observable product of more nuanced underlying linguistic strategies being employed more frequently by women. In particular, they use email greetings, along with other non-transactional elements, to negotiate social and professional relationships with co-workers, both male and female, before asking them to do something. They do this in a way that is not found in the emails written by male employees, and as such a richer and sociolinguistically explainable distinction emerges between the gender that is not achieved by relying on the relative frequencies of *hi* alone.

As noted above, in contrast to females, male employees more frequently use the *hi* + name greeting in information-giving emails than those including requests or directives. That said, even though the *hi* + name greeting is most commonly (56.21%) used by female authors when making requests for information or giving directives, the remaining 43.79% are found in emails which are information-giving:

Example 101

<From: kay.mann@enron.com>
<To: dale.rasmussen@enron.com>
<Subject: ABB Transformer Purchase and Sale Agreement>

Hi Dale,
I thought I'd send this along. It is the first draft of the 'break out' contract for the ABB transformers which go with the LM's. I thought Sheila told me that your stuff is in immediate need to getting this and the LM break out in shape. This has not been reviewed with ABB yet (in fact, I haven't read it yet), but just in case you want to give it a look, I thought I would pass it along.
Kay

Example 102

<From: kim.ward@enron.com>
<To: david.hensel@enron.com>
<Subject: FW: Request for electricity proposal>

Hi David,
I got your name from Don Black as the replacement for Roger Ponce. I am on the West Gas Origination desk for ENA in Houston and often run across customers looking for gas supply, particularly in California, that would be considered retail customers. I am forwarding a request I received this morning that EES might be interested in. I will continue to forward such information to you unless you let me know you are not the person I should be contacting.
Thanks,
Kim Ward

This greeting is used in this way almost as frequently by women as it is with requests and directives. Although much of the same informality and solidarity work is being done through its use in instances such as in Example 101 and 102, it is not used to reduce the power of a request or directive. Recent studies (Bamman 2014; Brezina and Meyerhoff 2014) have argued that research which aims to draw distinctions between groups of language users overlooks the importance of an individuals' linguistic variation within those groups. If analysis moves beyond group-level distinctions here to an idiographic examination of individuals who make up the groups, different patterns of preference emerge from individual to individual (Table 28). The first point to note is that 21 of the 34 females in EEC80 use *hi* + name, only 13 of whom use it more than once. These 13 female authors are from the full range of hierarchical positions, from Vice Presidents (Sager and Shackleton) to assistants (e.g. Fleming, Phillips), and the contexts in which these different occupational groups use *hi* + name differs. Those in higher positions, (Mann, Heard,

Table 28. Emails containing *hi* + name in female EEC80 data

		Hi + name	**Requests/Directives**	**Informational**
mann-k	Lawyer	242	142	100
heard-m	Lawyer	81	51	30
sager-e	Vice President	52	21	31
shackleton-s	Vice President	21	15	6
scott-s	Ana, Spec, Ass	17	12	5
fleming-r	Assistant	16	0	16
phillips-c	Assistant	14	5	9
watson-k	Director/MD	14	12	2
cash-m	Director/MD	10	6	4
causholli-m	Ana, Spec, Ass	6	2	4
taylor-l	Assistant	6	4	2
rodrique-r	Ana, Spec, Ass	2	0	2
symes-k	Trader	2	1	1
beck-s	Pres, CEO, COO	1	1	0
kuykendall-t	Trader	1	1	0
lokay-m	Assistant	1	1	0
mcvicker-m	Assistant	1	0	1
perlingiere-d	Lawyer	1	1	0
scholtes-d	Trader	1	0	1
thompson-p	Assistant	1	1	0
ward-k	Director/MD	1	0	1
Total		**491**	**276**	**215**

Shackleton, Watson and Cash), with the exception of Sager, more frequently use this greeting in emails expressing requests and directives. In contrast, those lower down the hierarchy (Fleming, Phillips, Causholli and Rodrique) primarily reserve its use for emails which give information to their recipients. In fact, Rosalee Fleming, assistant to Kenneth Lay (Enron founder, Chairman and CEO) uses the greeting exclusively in emails giving information (usually when passing information on for Lay) (Example 103 and 104). This is indicative of Fleming's role in the company and is one that is shared with Cathy Phillips, assistant of Mike McConnell (Enron CEO), who also uses *hi* + name greeting with a greater number of information-giving emails (Example 105).

Example 103

<From: rosalee.fleming@enron.com>
<To: sherri.sera@enron.com>
<Subject: Re: Jeff Skilling does MS150>

Hi Sherri –
Ken said that he will pledge $750.00 - $5 per mile. Sally will pass on to Holly in the Foundation office to prepare a check to get to you.
Rosie

Example 104

<From: rosalee.fleming@enron.com>
<To: john.hardy@enron.com>
<Subject: Re: Prospective Meeting with Jim Harmon, Chairman of EXIMBank,>
<Cc: joseph.sutton@enron.com, cindy.adams@enron.com>
<Bcc: joseph.sutton@enron.com, cindy.adams@enron.com>

Hi John –
Ken said that given he will be in town on March 15, he will plan to attend the meeting also. By way of this e-mail, I'll check with Joe's office to see what time it's scheduled.
Thanks.

Example 105

<From: cathy.phillips@enron.com>
<To: john.ambler@enron.com>
<Subject: Re: Senator Gramm at Houston World Affairs Council>

Hi John –
I thought I would let you know that Mike McConnell will be in London on business on Friday and therefore will be unable to attend.
Thank you.
Cathy Phillips

Besides this inter-group variation across occupations, there are individual author differences within these groups. For example, Marie Heard and Kay Mann are both lawyers, they are the two most frequent users of *Hi + name,* and they both show a preference for using this greeting in emails that give requests or directives. In terms of results thus far, they are virtually indistinguishable. However, the way in which they use *hi*, both in terms of form and function, is different. First, Heard is consistent in her use of the recipient's name after *hi*, as every instance of *hi* in her emails is followed by the recipient's name(s). This is in comparison with only 242 (76.34%) of Mann's. Second, in her 242 instances of *hi* + name, Mann is strikingly uniform in the form of the greeting, following the recipient's name with a comma in 239 (98.76%) of cases (Figure 34).

Figure 34. 10 of 239 concordance lines for *Hi* + name, in Kay Mann's emails

N	Concordance
1	services? Thanks, Kay <Message-ID: > Hi Rose, I hate to be a pain, but we are trying to
2	essage-ID: > fyi <Message-ID: > Hi Kathleen, Here's another deal like CA
3	me think about it. <Message-ID: > Agreement] Hi Reagan, Have you revised the exhibits to clarify
4	I get the latest version? Kay <Message-ID: > Hi Steve, We need wiring instructions for the
5	a lock installed on her door. <Message-ID: > Hi Lee, Just wanted to check to see if you will be
6	me. Enron North America Corp. <Message-ID: > Hi Warren, I'm going to forward a couple of emails
7	Thanks for the reminder. Kay <Message-ID: > Hi Ed, Here's a certificate of incumbancy we need
8	Barbara Gray Jeff Hodge <Message-ID: > Hi Warrren, Could you please print all of these
9	for the upgrade. Thanks, Kay <Message-ID: > Hi Taffy, Did the Pompano Beach meeting get
10	and Herman ok with this? Kay <Message-ID: > Hi Herman, This is another turbine transaction. I

In contrast, Heard never follows the recipient's name with a comma in this way. Instead, she uses *hi,* + name followed by an exclamation mark in 52 (64.2%) of her 81 instances and *hi,* + name followed by a colon in 29 (35.8%) (Figure 35).

Figure 35. 19 of 81 concordance lines for *Hi* + name in Marie Heard's emails

N	Concordance
1	ISDA initialled by you. Have you <Message-ID: > Hi, Tracy! I have never received the signature pages for the
2	it wait until Sara returns? Thanks! Marie <Message-ID: > Hi, Gordon! This is a Chilean counterparty that Nidia
3	we get a worksheet. Thanks! Marie <Message-ID: > Hi, Robbi! How are things up your way? It's incredibly
4	that I passed along to the attorney. <Message-ID: > Hi, Tanya! Here is CSFB's draft Schedule and Paragraph
5	if you need anything else. Marie <Message-ID: > <To: > Hi, Lech! My understanding is that you all will be handling
6	anything else. Thanks! Marie x33907 <Message-ID: > Hi, Mark! Francisco asked me to forward this e-mail to
7	forwarded on to you to assign. Marie <Message-ID: > Hi, Joe! Per my voicemail, will you send me the two
8	the heck, here they are! Thanks! Marie <Message-ID: > Hi, Veronica! Have you had a chance to review their
9	> Thanks, that's nice to hear! <Message-ID: > Hi, Tracy! I couldn't remember if I was supposed to send
10	is best for you? Talk to you later. Mare <Message-ID: > Hi, Greg! Wendy Conwell in Credit here in Houston
1	<Message-ID: > Hi, John: Can you answer Athena's questions re Navajo?
2	is for Elizabeth's project. Thanks! Marie <Message-ID: > Hi, Kim: Can you send me a representative copy of a
3	<Message-ID: > Hi, Dianne: Thanks for responding so quickly. It seems
4	I'll arrange for pickup. Thanks! Marie <Message-ID: > Hi, Cindy: Enron Corp. executed a guaranty dated as of
5	<Message-ID: > Hi, Patrick: Can you provide me with the date for Deal
6	<Message-ID: > Hi, Maribel: I don't know if you were copied on Tana
7	<Message-ID: > Hi, Georgi: We have an entry in our database that states
8	<Message-ID: > Hi, Brian: The ISDA is being faxed to you now. Marie
9	(713) 853-3907 Fax: (713) 646-3490 <Message-ID: > Hi, Steve: What is your exact location so that hopefully,

On the one hand, Mann consistently uses the same *hi* + name with comma greeting with all participants and in emails with various purposes. Heard, on the other hand, appears to adapt her greeting form in relation to who she is writing to and what she is writing about. Her correspondence with one particular colleague Robbi Rossi (robbi.rossi@enron.com) is especially intriguing in this regard. Heard emails Rossi a total of 33 times in her dataset, the majority of which are work-related, for example:

Example 106

<From: marie.heard@enron.com>
<To: robbi.rossi@enron.com>
<Subject: FW: Uecomm - Enron Master Agreement>

Robbi:
I left you a voice mail on Friday. I looked at the UEComm Master and had some comments--such as our name is wrong, the cross default threshold for us should be US $ and not AUD, and there were a few other questions I had.
Marie

Example 107

<From: marie.heard@enron.com>
<To: robbi.rossi@enron.com>
<Subject: RE:>

Sooner than January? I'll let you know if I hear of anything. I'm sure Mark H. is planning on you just integrating into ENA's legal group.

In these work-related examples, Heard does not use *hi* in her greeting, or disposes of the greeting altogether. However, on 22nd August 2001 Heard initiates a conversation thread with Rossi that is partially work-related but also in which she approaches Rossi to assist her in finding some fabric to cover furniture while she has her home decorated (Example 108). In this email, she employs the friendly *hi* + name! greeting, with the exclamation mark contributing to this informality. Again, as above, this greeting is used in combination with other relational features, particularly negative politeness. Heard expresses her wish to not interrupt Rossi, and engages in a phatic exchange (*how are things going now*...). After the request for assistance with the fabric is made, she shows she is accommodating to Rossi's plans (*we didn't know how that would fit into your plans*). The use of the informal greeting along with these other politeness strategies may be employed by Heard given that this is an unusual request, outside of the boundaries or norms of their regular workplace interaction.

Example 108

<Date: Wed, 22 Aug 2001 14:12:57 -0700 (PDT)>
<From: marie.heard@enron.com>
<To: robbi.rossi@enron.com>
<Subject: Decorator>

Hi, Robbi!
Thought I'd e-mail you rather than call and interrupt you. How are things going now that Lance and Carol are there? Hopefully, better. Mark gave me my review Thursday and I kept a copy of it to show you. Kristina managed to get her digs in, but, fortunately, Mark saw right through them.
Our contractor is going to start work on our house the week of Labor Day, so we are having to pack as much as we can to move upstairs. We thought it would be perfect if we could find fabric to recover our furniture so that maybe we could send it out while the construction is going on. If your offer's still good, we would like your assistance in looking for fabric. We were hoping to take a day off to look for fabric but didn't know how that would fit into your plans. Would you let me know?
Also, we need to schedule a lunch with Erica and Jane. What is best for you?

Talk to you later.
Marie

The same features are found in an email further on in this conversation, in which Heard lists the fabrics she is interested in sampling. Here, the greeting is followed by a phatic exchange about time off work, as well as the informal *thanks!* (Example 109).

Example 109

<Date: Wed, 5 Sep 2001 09:22:10 -0700 (PDT)>
<From: marie.heard@enron.com>
<To: robbi.rossi@enron.com>
<Subject: Fabric>

Hi, Robbi!
Hope you are enjoying your time off. I know I certainly enjoyed Friday and Tuesday and hated to come back today.
Anne and I went to the Decorative Center yesterday, and tried to check out some samples at Beacon Hill (Robert Allen Fabrics) and they couldn't find your account. Do you have an account there, or was I mistaken?
Anyway, I thought I'd e-mail you since I don't have your home number in case you were going to the Decorative Center later this week to see if you could check out some samples for us if you had time.

[...]

If you don't go back to DC or have already gone, we can maybe find the Robert Allen samples at Expo. They sell his fabrics there.
Thanks! We really appreciate your help.
Marie

Finally, as the 'fabric' conversation develops and draws to an end, Heard continues to use the *hi,* + name! greeting:

Example 110

<Date: Fri, 19 Oct 2001 07:46:31 -0700 (PDT)>
<From: marie.heard@enron.com>
<To: robbi.rossi@enron.com>
<Subject:>

Hi, Robbi!
How are things going? Any more news?
We have finally decided on our fabric and the amount of yardage. What's the next step?
Thanks!
Marie

Example 111

<Date: Mon, 5 Nov 2001 12:08:18 -0800 (PST)>
<From: marie.heard@enron.com>
<To: robbi.rossi@enron.com>
<Subject:>

Hi, Robbi!
How are things up your way? It's incredibly busy here--entering into numerous master netting agreements. Hope things quiet down soon. I can't take much more of this turmoil!
Anne got her fabric from Schumacher Saturday. Do you have an invoice or will you let us know the total when you get one? There wasn't an invoice included in the package. The toilet is so pretty.

Robbi Rossi is the addressee with whom Marie Heard uses the *hi + name!* greeting the most (five times), and this conversation about fabric contains all of these five instances. As has been shown, *hi* is used here with a pragmatic function, as Heard shows sensitivity to author and topic of the email, marking a section of non-Enron related communication apart from her regular workplace emails with Rossi. By looking at the 'Subject' field in the metadata for each of these emails, the lack of 'RE: indicates that in each of these emails, Heard is initiating the communication, rather than responding to Rossi. Furthermore, an examination of Rossi' responses shows that he does not use *hi* back to Marie, instead favouring a greeting of only her name:

Example 112

<Date: Fri, 19 Oct 2001 10:57:00 -0800 (PST)>
<From: robbi.rossi@enron.com >
<To: marie.heard@enron.com>
<Subject: RE:>

Marie,
It's been slow - but I think its going to pick up. Send me a list of what you want to purchase and from where - as well as the address where you want it shipped. I will then firnd out pricing for you. It may take a little longer with the Robert Allen stuff since I have to open an account there. I do not anticipate any problems getting the account open. [...]

The fact that Marie Heard alters her use of *hi* + name in relation to the communicative situation, addressee and function of the email distinguishes her linguistic behaviour from that of Kay Mann. Mann consistently uses *Hi + name,* with a comma, indicating that she does not adapt her use of this greeting with the same pragmatic motivation as Heard does. As such, a distinction is drawn between these two authors that could not be done on the basis of nomothetic quantitative generalisations. Identifying behavioural patterns such as these highlights the value of an idiographic approach to forensic author profiling. The brief idiographic comparison of Heard and Mann is not only concerned with how frequently they use a certain feature, but also reveals their linguistic behaviour with this feature across contexts and audiences. It provides the analyst with an idea of the kinds of communicants they are, or 'what kind of linguistic person(s)' (Grant 2008: 223) they are. Heard is someone who uses *hi* in different forms with different pragmatic functions across different contexts. Specifically, she modifies the punctuation which she attaches to *hi + name* greetings in a way that responds to the communicative situation in which she is writing. In contrast, Mann is far more uniform in her use of the form *hi + name* followed by a comma, which she uses consistently.

To draw the comparison between linguistic profiling and offender profiling, Canter and Youngs (2009: 151) explain that one of the main questions investigative psychologists face in a case is the question of consistency: how consistent is an offender's criminal behaviour with their behaviour in other aspects of their life? The author profiling problem can be mapped onto this in an authorship case: how consistent is a suspect's linguistic behaviour in any disputed text (one of which they are suspected of being the author) with their linguistic behaviour in their known (non-criminal) texts? Of course, this is the central issue in all authorship attribution

tasks. However, in a forensic situation, 'disputed' and 'known' texts can be compared on a dimension beyond the similarities and differences in linguistic features within them. The known texts of a suspect can be analysed in the same way as with Heard and Mann above, to gain a sense of what kind of communicants they are. In turn, the questioned text(s) can be examined in terms of whether the linguistic behaviour exhibited in it (or them) shows any salient choices made in relation to, for example, the addressee of the text (a man/woman, a child, a vulnerable person, a person they know) and the purpose of the text (to threaten, scare, extort, exercise power over). Such information will provide answers to the question: what kind of communicant is this person? Given the data, the findings of the analyses can be compared, and a judgement can be arrived at as to whether the criminal linguistic behaviour in the questioned text is consistent with the non-criminal linguistic behaviour in the suspect's known texts.

Such an argument brings to the fore where the emphasis should be in 'author profiling'. To date, author profiling research has been nomothetic, focusing on quantitative findings in assigning a social profile to the author of a questioned document: what gender or age they are, what their native language is or what type of personality they have. However, the findings here suggest that analysts may find more value in building an idea of the *linguistic* profile of authors, both in questioned and known documents. Such an idiographic approach has the advantage of analysing an 'actual offender who exists in the real world', as Turvey (2012: 122) argues of idiographic offender profiling.

7.4 *why* and its reflection of author identity

The analysis of *for* found that linguistic choices are made by authors drawing on resources related to both gender and occupation. The investigation of *hi* identified that individual users use it with particular pragmatic functions and to particular participants, and it is this communicative behaviour that determines its use, rather than being bound to a particular gender or particular occupation. Whereas the Mann-Whitney U and Kruskal-Wallis tests performed at the start of this chapter found the relative frequencies of *for* and *hi* to distinguish between the gender only, they identified *why* as being statistically significant in distinguishing between authors in terms of both gender *and* occupation in EEC80. It is a significant feature for male writers, while at the same time being significant for traders. This was found

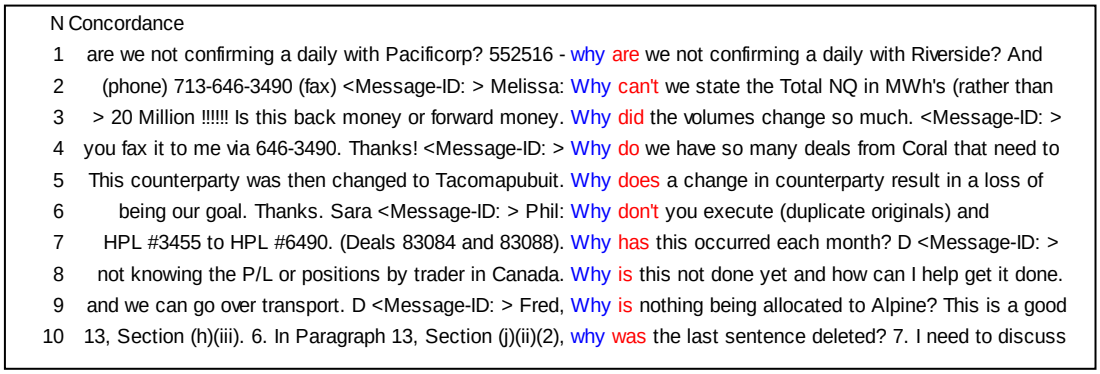
alongside results for other *wh-* adverbs, namely that *where* is significant for Presidents, CEOs and COOs and *when* is significant for managers. It was speculatively argued that such results may be indicative of different emphases for different occupation groups in Enron; whereas traders are concerned with *reason*, Presidents, CEOs and COOs are concerned with place while managers are concerned with time. The focus of this analysis is on *why*, and it reveals the different ways in which its use is determined, and sanctioned, by the individual's professional identity as well as the social relationships they have with their recipients.

By using ranks rather than values themselves, the Mann-Whitney U and Kruskal-Wallis tests take account of intra-group variation (Brezina and Meyerhoff 2014: 1), and it is on these ranks (explained in Section 4.4.2) that the statistically significant differences are based. Comparing relative frequency values themselves, the differences between the groups seem smaller. *Why* accounts for 439 (0.04%) of the 1,181,409 tokens in the male corpus, and 346 (0.03%) of the 996,796 tokens in the female corpus, and there are not huge differences in the relative frequencies across the occupation groups (Table 29). There are a total of 785 instances of *why* in the EEC80. These 785 instances are divided almost equally between interrogative adverbs (n=378, 48.15%) (Figure 36) and non-interrogatives (n=407, 51.87%) (Figure 37). These two types of *why* instances were distinguished in this corpus on the basis that in interrogatives, *why* precedes the verb.

Table 29. *why* frequency across occupations in EEC80

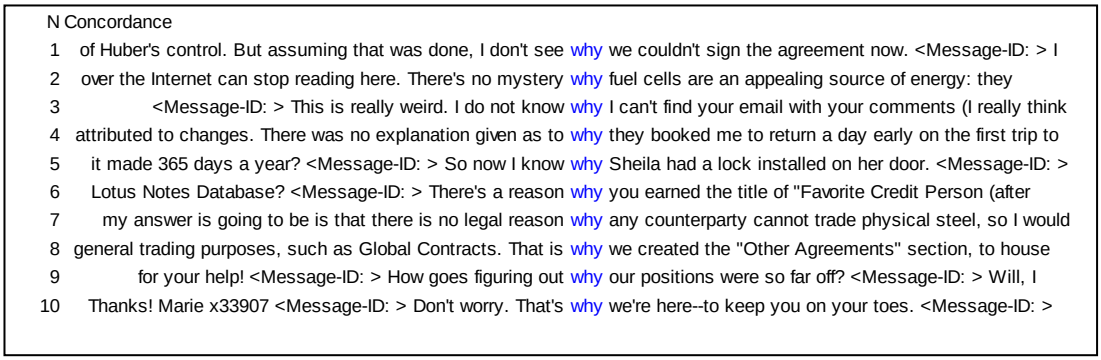
	Total tokens	***why***	**% of total tokens**
Pres, CEOs, COOs	266,500	94	(0.035%)
Vice Presidents	435,874	171	(0.039%)
Directors and MDs	407,047	130	(0.032%)
Lawyers	407,646	117	(0.029%)
Managers	197,518	57	(0.029%)
Traders	207,321	**109**	**(0.053%)**
Ana, spec, ass	225,365	103	(0.046%)
Assistants	30,934	4	(0.013%)

Figure 36. 10 of 378 Concordance lines showing *why* as an interrogative adverb in EEC80



N	Concordance
1	are we not confirming a daily with Pacificorp? 552516 - why are we not confirming a daily with Riverside? And
2	(phone) 713-646-3490 (fax) <Message-ID: > Melissa: Why can't we state the Total NQ in MWh's (rather than
3	> 20 Million !!!!!! Is this back money or forward money. Why did the volumes change so much. <Message-ID: >
4	you fax it to me via 646-3490. Thanks! <Message-ID: > Why do we have so many deals from Coral that need to
5	This counterparty was then changed to Tacomapubuit. Why does a change in counterparty result in a loss of
6	being our goal. Thanks. Sara <Message-ID: > Phil: Why don't you execute (duplicate originals) and
7	HPL #3455 to HPL #6490. (Deals 83084 and 83088). Why has this occurred each month? D <Message-ID: >
8	not knowing the P/L or positions by trader in Canada. Why is this not done yet and how can I help get it done.
9	and we can go over transport. D <Message-ID: > Fred, Why is nothing being allocated to Alpine? This is a good
10	13, Section (h)(iii). 6. In Paragraph 13, Section (j)(ii)(2), why was the last sentence deleted? 7. I need to discuss

Figure 37. 10 of 407 concordance lines showing *why* as a non-interrogative in EEC80



N	Concordance
1	of Huber's control. But assuming that was done, I don't see why we couldn't sign the agreement now. <Message-ID: > I
2	over the Internet can stop reading here. There's no mystery why fuel cells are an appealing source of energy: they
3	<Message-ID: > This is really weird. I do not know why I can't find your email with your comments (I really think
4	attributed to changes. There was no explanation given as to why they booked me to return a day early on the first trip to
5	it made 365 days a year? <Message-ID: > So now I know why Sheila had a lock installed on her door. <Message-ID: >
6	Lotus Notes Database? <Message-ID: > There's a reason why you earned the title of "Favorite Credit Person (after
7	my answer is going to be is that there is no legal reason why any counterparty cannot trade physical steel, so I would
8	general trading purposes, such as Global Contracts. That is why we created the "Other Agreements" section, to house
9	for your help! <Message-ID: > How goes figuring out why our positions were so far off? <Message-ID: > Will, I
10	Thanks! Marie x33907 <Message-ID: > Don't worry. That's why we're here--to keep you on your toes. <Message-ID: >

In terms of gender, males use a greater proportion of interrogative *why*s than females. The 439 instances of *why* in the male writing comprise 50.34% (221/439) interrogative and 49.66% (218/439) non-interrogative. In contrast, only 37.28% (129/346) of females' *why* instances are interrogative, with non-interrogatives accounting for 62.72% (217/346).

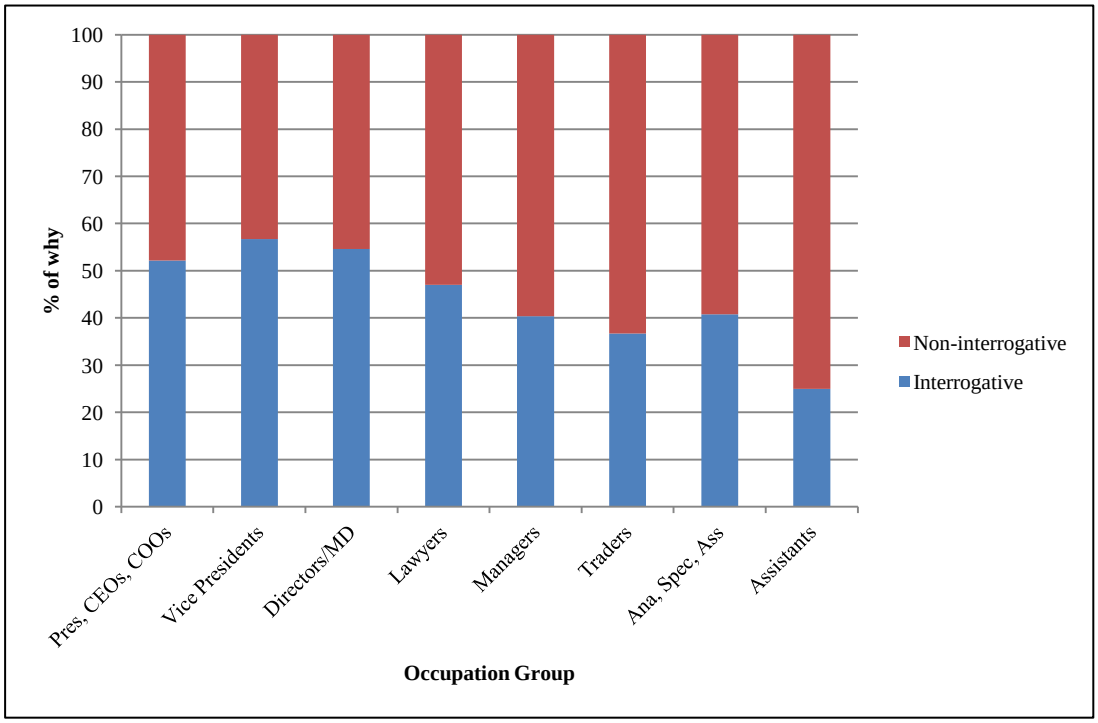
Although slightly less frequent in the EEC80 corpus, focus here is on the interrogative use of *why*, given the importance of questions in institutional and workplace interaction (Freed and Ehrlich 2010: 3). Tracy and Robles (2009: 131) argue that questioning is '*the* central communicative practice of institutional encounters', and that it enacts and reflects 'professional and lay identities of key parties'. Heritage (2004: 237) emphasises that 'there is a direct relationship between institutional roles and tasks on the one hand, and discursive rights and obligations on the other'. In the context of the workplace, Vine (2004: 27, 42) identifies interrogatives as a 'control act', or 'a way of getting someone to do something'. In asymmetrical power relationships in a workplace, the more powerful interactants are

at liberty to ask questions and, as Wang (2006: 529) argues, 'questions are a possible means for dominant participants to exert power over subordinate individuals'. She continues:

> some participants dominate the position of asking questions, while the other participants have very limited opportunity to ask questions or are deprived of the opportunity to ask questions due to different institutional identities.
>
> (Wang 2006: 540)

The asking (or not) of questions is a linguistic act closely related to the institutional identity of the author in question, primarily reserved for more powerful participants, This suggests that in EEC80 the use of *why* may be more closely linked to occupational group rather than truly distinguishing between gender. This argument is further supported when the use of *why* as an interrogative is compared across the eight occupation groups (Figure 38). In the EEC80 sample, *why* is used more frequently as an interrogative adverb by the three occupation groups highest up the hierarchy: Presidents, CEOs and COOs, Vice Presidents and directors/managing directors. In contrast, those lower down the hierarchy—lawyers and below—use *why* more frequently as a non-interrogative. Therefore, the function and use of *why* by authors appears to be closely related to their institutional identity.

Figure 38. The use of *why* as an interrogative adverb across the EEC80 sample



As such, it might be argued that it is this element of the author's identity which is the overriding factor in its use, and it is returned as a significant feature for males because more males populate higher ranking Enron positions than females. Furthermore, traders, the group for which *why* is statistically significant in the first instance, can be distinguished from other occupation groups on the basis that they use it as an interrogative less frequently (36.70%) than any of the other groups (with the exception of assistants, who use *why* four times in total).

This kind of stylistic distinction enriches the statistical results, and provides a sociolinguistic explanation for the findings of a nomothetic approach. However, a further stylistic analysis of different traders reveals that *why* is used with different functions and different frequencies even within this group. One trader, Don Baughman, does not use *why* at all in his data, and so the belief that *why* is characteristically a male and trader feature is not accurate for him. Therefore, he exemplifies the criticism of generalised nomothetic results in both psychological and author profiling that any 'real' individual in a given case can be an outlier, to whom the generalisation being made does not apply (Coulthard et al. 2011: 538; Turvey 2012: 122). Baughman brings this drawback into sharp focus, as he defies the overall general nomothetic results. In a forensic context, if his emails constituted a set of 'disputed' documents, and the analyst attempted to determine the gender and or occupation of the writer using the nomothetic quantitative results obtained above, then, given his non-existent use of *why*, it is unlikely (on the basis of this evidence alone) that he would be identified as a male trader.

Continuing with the remaining nine traders, some use *why* more frequently than others (Table 30).

Table 30. *why* across the ten EEC80 traders

	Tokens	**Total *why***		**Interrogative**		**Not Interrogative**	
Diana Scholtes	1,871	2	(0.11%)	0	(0.00%)	2	(100.00%)
Kate Symes	58,754	50	(0.09%)	11	(22.00%)	39	(78.00%)
Darron Giron	15,415	11	(0.07%)	1	(9.09%)	10	(90.91%)
Eric Bass	25,481	17	(0.07%)	11	(64.71%)	6	(35.29%)
Mark Guzman	13,227	8	(0.06%)	4	(50.00%)	4	(50.00%)
Susan Pereira	21,41	1	(0.05%)	1	(100.00%)	0	(0.00%)
Joe Parks	44,98	2	(0.04%)	1	(50.00%)	1	(50.00%)
Tori Kuykendall	47,35	2	(0.04%)	0	(0.00%)	2	(100.00%)
Chris Germany	77,597	16	(0.02%)	11	(68.75%)	5	(31.25%)
Don Baughman	3602	0	(0.00%)	0	(0.00%)	0	(0.00%)

Kate Symes and Chris Germany have the most tokens of all the traders. While *why* makes up 0.09% of Symes' total words, it accounts for only 0.02% of Germany's. Similarly, different individual traders use *why* for different things. Whereas authors such as Symes and Darron Giron use it predominantly as a non-interrogative, this preference is reversed for others such as Germany and Eric Bass. *Wh*- questions such as *why* are 'information-seeking' questions (Maley and Fahey 1991: 6). In some contexts, such as courtroom trial discourse, they have been considered as less controlling and coercive than other question types, such as yes/no or tag questions (Harris 1984: 10–11; Conley and O'Barr 1998: 24; Gibbons 2003: 102). However, in the workplace they are controlling as they 'seek to elicit completion of a proposition from an addressee', and 'require the addressee to introduce new factual material' (Wang 2006: 544). *Why* questions are particularly threatening to the addressee's negative face, given that the information being sought is a reason, a justification or an explanation for actions or events that have transpired or are transpiring. Brown and Levinson (1987: 71) identify that there are a number of factors influencing the 'seriousness' of a face-threat. In terms of *why* questions, two of the most important factors are the topic about which the question is made as well as the relationship between the asker and the recipient. This is another dimension through which individual traders' linguistic patterns can be distinguished from each other, and is demonstrated by an idiographic comparison of how interrogative *why* is used by Kate Symes and Eric Bass. Kate Symes more frequently

uses *why* as a non-interrogative (Table 30), and when she does use it in interrogatives they are all to do with work-related topics:

Example 113

<From: kate.symes@enron.com>
<To: evelyn.metoyer@enron.com>
<Subject: Re: Commission for Bloomberg>

Tom and Mark are not spot traders, they're cash traders. This is what I was referring to yesterday - that Real Time is now being charged $15 for trades - but cash and term traders should still be charged the traditional $.005. **Why is Bloomberg recognizing these trades as Real Time**?
Kate

Example 114

<From: kate.symes@enron.com>
<To: mark.guzman@enron.com>
<Subject: Los Angeles Department of Water and Power IS NOT LADWPPX>

Just FYI - I believe you entered these deals on Jan. 16. What were you thinking? How long have you worked here? Can you say, 'wrong counterparty name'? Also, **why don't you have your own trader ID yet?** Geez.
Sincerely,
Kate
P.S. Don't worry - I corrected them for you.

Example 115

<From: kate.symes@enron.com>
<To: kimberly.hundl@enron.com>
<Subject: Deal Changes - No Confirm>
<Cc: diana.scholtes@enron.com>
<Bcc: diana.scholtes@enron.com>

Kim -
I've finished changing the list of deals you faxed over to no confirm. Please let me know if you continue to see these deals on your new deal report. I still had questions on some of the Short Term and Long Term Northwest deals.
On those I will defer to Diana Scholtes. I also have questions on the following deals:

557772 - **why are we not confirming a daily with Pacificorp?**
552516 - **why are we not confirming a daily with Riverside?**

In these examples, Symes is seeking information from her recipients Evelyn Metoyer, Mark Guzman and Kimberly Hundl regarding trading issues. Here, *why* questions are being used as Symes exercises and projects her professional identity in a situation in which she has the hierarchical superiority and discursive rights to do

so. The *why* questions may still be perceived as attacking the negative face of the recipients; they impede and question their actions or the actions of their team. However, this face-threat is mitigated because the communicative situation is such that the asking of *why* questions by Symes is sanctioned.

In contrast, none of Eric Bass' eleven *why* questions relate to trading. Rather, they are social in nature:

Example 116

<From: eric.bass@enron.com>
<To: timothy.blanchard@enron.com>
<Subject: Re: CWS>

you know that all of us (Matt, Chad, and I)are just fucking with you about the EES bullshit. [...] i know that you, matt, and chad think that i had nerves and that is why i didn't take the test, but **why the hell would i have nerves about a test that has no bearing whatsoever on my future.** so quit the bullshit, and, TEXAS will still beat the shit out of LSU on Sat.

Example 117

<From: eric.bass@enron.com>
<To: shanna.husser@enron.com>
<Subject: Re: Bryan Hull is moving on. Let's celebrate !!!>

why the hell were you invited?

Example 118

<From: eric.bass@enron.com>
<To: hector.campos@enron.com>
<Subject: Re:>
<Cc: brian.hoskins@enron.com>

why not bitch?

In Example 117, Bass is emailing Shanna Husser, asking why she has been invited to Bryan Hull's leaving celebration (mentioned in the subject). A Google search for 'Shanna Husser Enron' returns a Facebook profile for a Shanna Husser Bass, who worked at Enron and lives in Houston Texas. Based on this, it appears as though Eric Bass and Shanna Husser are now husband and wife. Other emails from Bass to Husser include *you can cook this for me, by the way dinner better be good tonight,* and *whatever you think babe – i will trust you to make plans.* While at first glance his question *why the hell were you invited?* seems a 'bald on record' face attack (Brown and Levinson 1987: 69), in the context of Bass' other emails to Husser (some 114) it is clear that they have a good personal relationship. This personal

relationship between participants dilutes the face threat, as it is more likely to be (interpreted as) a joke than a genuine face attack. Similarly, Bass has a close personal relationship with Hector Campos, the recipient of his *why not bitch?* email (and Brian Hoskins CCd). 'Subject' fields of other emails between Bass and Campos are 'Dave Chappelle' (a famous American comedian), 'Happy Hour' and 'Sherlock's on Friday Night', all of which involve the planning of social events. The *why not bitch?* question is in an email conversation about a paintballing trip. The content of the conversation is as follows:

```
Hector Campos:  I'm not going to make it to paintball on saturday.
                sorry. please take me off the list. -Hector
    Eric Bass:  why not bitch?
Hector Campos:  I don't want to go
    Eric Bass:  pussy
```

As with Shanna Husser, Eric Bass has a strong personal relationship with Hector Campos outside of the workplace. This relationship is also reflected in the use of curse words *bitch* and *pussy* as terms of endearment (Jay 1992: 177). Therefore, in contrast to Kate Symes, for whom it was her professional identity which permitted her use of *why* questions in work-related emails, it is Eric Bass' close personal relationship with his recipients which influences his use of *why* interrogatives.

The examination of the use of *why* in relation to function and addressee has identified that, to conclude that it is a significant feature for men and for traders is a gross over-simplification of the linguistic reality. As with the other two features (*for* and *hi*), the nomothetic results represent only the surface of this reality, the bare minimum in terms of what can be said about *why* in relation to gender and occupation: that men and traders use it the most. Differences in the frequency of use of all three of these words are only indicative of larger scale, sociolinguistically and pragmatically explainable, underlying differences in communicative preferences and patterns of behaviour. The identification and analysis of such behaviour can be of value to forensic linguists who are attempting to gain an understanding and insight into the linguistic practices of a particular group of language users or identify what kind of person wrote a text. The nomothetic approach used to predict social characteristics of unknown authors, it has been found, can be inaccurate. While statistical results may be important, the identification of quantitatively salient

linguistic features should serve only as a first step in the analysis. In turn, a more stylistic, idiographic approach could be taken in the analysis of the individual. Rather than simply counting the frequencies with which they use particular features, their communicative behaviour in relation to context, addressee and purpose could be examined, to explore whether these linguistic preferences reveal different elements of the multidimensional identities which they are projecting, and to answer the question, *which kind of linguistic person wrote this text?*

7.5 Chapter conclusion: what next for author profiling?

In contrast to the problem of authorship *attribution*, authorship *profiling* has received much less research attention. With the advent of larger datasets (e.g. 'Big Data'), more powerful computers, and increased impetus from domains such as law enforcement, marketing and advertising, interest and momentum has rapidly built in the processes by which social characteristics of (unknown) writers can be identified. The work that has been done over the last ten to fifteen years in quantitative author profiling has focused on categorising documents on the basis of authors' social variables. Such studies have invariably reported good results, across a range of text types and a range of social traits. These studies (e.g. Argamon et al. 2003; 2009; 2013; Noecker et al. 2013; van Halteren et al. 2008; Koppel et al 2005; Estival et al. 2007) identify linguistic features that statistically significantly discriminate between authors from different social categories—whether that be age, gender, level of education, native language, or personality type—and then often use these features to predict which categories apply to the author of an anonymised text.

This chapter set out to critique and develop this process, both methodologically and theoretically. It identified features which statistically distinguish between male and female authors, and employees with different occupations in the EEC80 sample, and analysed how these words are used across various contexts by various groups and individual group members. The first part of the analysis continued in the tradition of quantitative or 'nomothetic' approaches to author profiling, in that statistical techniques were used to identify features which are used with significantly different frequencies by different groups of authors. The findings problematised this approach to author profiling in three main ways. First, the majority of the 291 features included in the statistical analysis discriminated neither between gender nor occupations. This suggested that general, quantifiable

differences across groups are not as abundant as author profilers might hope. Second, those features which did distinguish between the genders were inconsistent with those emerging from an already inconsistent field of existing research and findings. Third, many of the same features were found to discriminate between both gender and occupation. Based on these results it was argued that quantitative nomothetic approaches to author profiling take an oversimplified one-dimensional view of identity. In such a view, individual aspects of a person's identity are considered as separate from each other, and separate from the context in which the language is produced. In turn, the stylistic analysis of three words—*for*, *hi* and *why*—found that the use of these words is determined by underlying linguistic and pragmatic preferences and patterns. Factors such as the function of the communication in question, the relationship between author and recipient, and the complex projection and negotiation of social and professional identities all influence their use. Additionally, the in-depth study of individuals, or an 'idiographic' approach, found that social groups are not homogenous in their language use, with different authors drawing on different social and contextual resources in their use of the same linguistic features.

On the basis of the findings here, three main points can be made about author profiling, and possible directions for the future. First, it might be that nomothetic, stylometric approaches to authorship profiling are over-reliant on statistical significance. All such methods begin by identifying features which statistically discriminate between groups. Understandably, for the sake of time and space constraints in research, the threshold of what is important and what is not in a particular study needs to be drawn somewhere, and statistics have a lot to offer both linguistics (e.g. Oakes 1998; Gries 2009) and forensic science (Lucy 2005). However, the concern is that in practice, such a pre-occupation with statistical significance may result in important linguistic differences going unnoticed. With *for* and *hi* for example, no significant difference was found in frequency of use across occupation. However, the subsequent stylistic analysis suggested that professional identity was the major influence in the use of these features, rather than gender. A debate around *p* values and statistical significance frequently recurs in the natural sciences. A 2014 issue of the journal *Ecology* (Volume 95, Issue 3), for example, hosted a forum in which the advantages and disadvantages of purely statistical approaches were discussed. While reliability is of utmost importance in the field of

forensic linguistics, it may be worthwhile to have a similar debate in relation to authorship profiling. A potential location for this debate is as a special issue of the forensic linguistics journal *Language and Law (Linguagem e Direito)*, which would bring together articles and commentaries from statistical and non-statistical authorship analysts.

Second, criminal/offender profiling in forensic psychology has been criticised for being based on outdated theories of personality and behaviour (e.g. Snook et al. 2008: 1259). In the same way, algorithmic approaches to author profiling are guilty of being based on outdated theories of identity. Sociolinguists are increasingly eschewing 'unhelpful static universalities about how all women, all English speakers or all old people behave', and are moving towards an 'interactionally based conceptualisation' of identity (Angouri and Marra 2011: 1). Approaches to author profiling which are sensitive to this more complex view of author identity may be more helpful than those which correlate specific linguistic features with rigid social categories and produce results that do not accurately represent any real language user (e.g. Don Baughman and his lack of *why*). Importantly, author profiling should avoid exploiting social traits as determinants of language use, and instead adopt the theory that 'identity is something we actively do, rather than something we passively are' (Marra and Angouri 2011: 1).

Third, and related to the two previous points, is the argument that author profiling should move beyond the simple presentation of linguistic features which appear to discriminate between any groups under analysis. This should be supplemented by the analysis of contextually-sensitive patterns of linguistic and communicative behaviour within groups and individuals. As a field, profilers may learn more about authors and their groups if the focus is on the pragmatic and interactional behaviours of authors rather than their use, more or less, of an arbitrary set of linguistic features within a specific text type. Of course, this is a more difficult prospect for the forensic linguist. Tasked with identifying the kind of person who wrote a given text, the analyst needs to use all of the resources at their disposal. Therefore, the comparison of the features exhibited in the text with those apparently 'typical' of certain social groups is naturally going to be an attractive first stage. However, this chapter has shown the difficulty of mapping individuals onto abstract group-level generalisations, and at the same time highlighted the value of analysing the individual authors who constitute these groups, and how idiographic analyses

can aid forensic investigation. In comparison with authorship attribution, author profiling research is in its infancy, and computationalists, corpus-, socio- and forensic linguists should collaborate at this early stage to develop a combined approach. This may help avoid the situation currently unfolding in the authorship attribution community wherein linguists and computationalists are working largely independently, developing divergent and competing methodologies.

8 Conclusions, contributions and future directions

8.1 A corpus linguistic approach to authorship analysis

The central aim of this study was to investigate how a corpus linguistic methodology can be used to address both the theoretical and methodological challenges in the field of forensic authorship analysis. Chapter 2 discussed, in detail, the nature of these challenges, and they are worth reiterating and summarising here, as the study concludes. The traditional theory of idiolect as everything that person *could* say or write in a given language is clearly too idealised and abstract to be of any practical use to the forensic linguist. Forensic linguists have recently begun reconceptualising the notion of idiolect, or 'idiolectal style' (Turell 2010: 217), as the measurable distinctiveness of linguistic features used by writers, when tested against the norms of the population from which they are taken. In other words, the 'population-level' distinctiveness of language use (Grant 2010: 515) in relation to the 'Base Rate Knowledge' for a given linguistic community (Turell 2010: 217; Turell and Gavaldà 2013: 499) can be considered as offering idiolectal evidence for an individual author.

In addition to this theoretical issue, the current methodological situation in authorship attribution research is one in which two diverging approaches have developed. On the one hand, there are stylistic approaches in which the analyst endeavours to manually identify potential linguistic style markers which offer clues as to common authorship (or not) of disputed documents, when compared with sets of known writings from each of the candidate authors. On the other hand, stylometric approaches to quantitatively measuring author style have attracted a lot of research attention. These approaches rely on comparing the relative frequency with which texts and authors use particular linguistic features, such as function words, parts of speech, or syntactic patterns, to algorithmically and automatically categorise texts with common authorship. Both approaches have their advantages and disadvantages. While the identification of style markers in stylistic analyses is grounded in theories of language variation, stylistic results have been criticised as being too subjective, intuitive and unreliable, as well as being impossible to generalise beyond the scope of the particular case in question. In contrast,

stylometric techniques offer a more objective, replicable and statistically reliable alternative, but the linguistic features they rely on for measuring similarity between authors are not underpinned or motivated by a theory of idiolect, and the statistical results they return are rarely, if ever, supplemented by linguistic or stylistic evidence. Therefore, it is often impossible to explain *why* a particular algorithm or set of features have worked in identifying authorship.

Finally, author profiling is in the process of developing as a field of research which is concerned with predicting as much a possible about the social characteristics of an author. All of the existing research in this field is stylometric in nature, and aims to correlate the relative frequency of linguistic features with particular aspects of individuals' identities, such as age, gender, level of education, native language and personality type. However, in practice, the kind of quantitative generalisations produced by these studies are such that they cannot be reliably applied to any one text or author. A parallel can be drawn between stylometric author profiling work and that of quantitative nomothetic approaches to criminal profiling in forensic psychology. While also proving very popular in forensic psychology, nomothetic approaches to profiling have been criticised as being too abstract, and not describing a real offender. Idiographic approaches have been proposed as a qualitative alternative, which involve the in-depth study of an individual, and therefore provide more 'concrete' evidence of their behaviour. It was argued that stylometric approaches to author profiling could be augmented and enhanced by similar qualitative linguistic evidence.

What has been shown throughout this thesis is that a corpus linguistic approach can make substantial steps in tackling these issues. As McEnery and Hardie (2012: 26) point out, the collection and analysis of large amounts of observational data has been used to create, accept and reject theoretical hypotheses across a range of natural sciences, including astronomy, geology, palaeontology and theoretical physics. In the same way, they continue, observation of language through corpora can be used to test linguistic theory. In turn, as was demonstrated in Chapter 5, a corpus linguistic methodology can be used to provide empirical evidence in support of the theoretical notion of idiolect. In methodological terms, corpus linguistics has traditionally offered the linguist the opportunity to combine quantitative and qualitative analyses, with the former providing 'statistically reliable and generalisable results' and the latter 'greater richness and precision' (McEnery

and Wilson 2001: 77). Such a combination has been adopted throughout all the analyses in this study. Overall, the analysis chapters in this thesis, summarised below, empirically tested Coulthard's (1994: 40) idea twenty years ago that corpus approaches offer the potential for improved methodologies in authorship analysis. The findings produced throughout this thesis support Coulthard's claim, and demonstrate the ways in which a corpus linguistic approach can help move the field beyond the current competitive methodological situation of *Stylistics versus Statistics*.

8.2 Summary of results

The aim of Chapter 5 was to empirically test the theory of idiolect, and in particular the belief held by corpus linguists and psycholinguists (e.g. Nattinger and DeCarrico 1992; Wray 2002; Schmitt et al. 2004; Hoey 2005; Barlow 2010; 2013) that collocation patterns and lexical co-selection preferences are personal, unique and idiolectal for language users. To do this, the Enron Email Corpus was used as a reference set, representing a Base Rate Knowledge of the Enron linguistic community, against which the population-level distinctiveness of word sequences could be measured. Overall, the results produced offered evidence to suggest that the collocation patterns and the sequential lexical co-selections authors make are often unique. In particular, focus was on three very common words in the corpus, *I, deal* and *please*, all of which, it was found, give rise to author-distinctive usage. The case of *I* highlighted that more can be learnt about authors' idiolectal stylistic language production by comparing the ways in which they use very frequent function words, than by simply comparing the relative frequencies with which they use them, as is common practice in stylometric work. With *deal*, a case was made for putting content words at the centre of discussions about idiolect. Content words are often avoided in authorship studies as they are too closely-related to the topic and register of the text, rather than being indicative of author style. While this is true if what is being considered is the relative frequencies of the individual words themselves, the results in this chapter revealed that, although writers share these words, the way in which they linguistically encode and collocationally package them can be unique. Finally, the markedly high frequency of *please* when compared with *The Corpus of Contemporary American English* identified *please*-mitigated directives as being one of the major speech acts in the corpus. A close analysis of a number of specific

directives, such as *please forward, please see, please put (x on my calendar/schedule)* and *please call/contact/give me a call*, revealed that when expressing the same speech act, with the same function and for the same purpose, authors produce different and distinctive linguistic output. Overall, the study of all three of these node words found that even with very commonly shared lexical items (*I* and *deal*), and speech acts (*please*-mitigated directives), authors' collocational preferences can be unique. The strength of this idiolectal evidence is enhanced when it is considered that these collocation preferences are distinctive when tested against a population of writers who are all writing within the same company, at the same time, using the same medium of communication and, often, writing about the same things with the same purpose. In turn, these distinctive collocational patterns can be explained by the fact that different authors have different linguistic experiences throughout their lives. They interact with different people in different situations, they have different family and friends, they read different books and they consume different types of media (social media, TV, film). As a result of this unique lifetime of different linguistic encounters and memories, authors have built and stored different associations between words, different lexical primings, and have formulated different word clusters for performing certain communicative roles. In turn, when they use language, these unique aspects of a person's cognition become identifiable manifestations of their idiolect. Therefore, based on the evidence of this chapter, and the argument that collocational patterns in an author's writing can be explained by a theory of idiolect, they provide useful features for the identification of authors in attribution tasks.

Chapter 6 built on the theoretical and idiolectal evidence in Chapter 5 and developed a method of authorship attribution which used word n-grams—strings of words between one and six in length—as a means by which to capture author-distinctive collocations and lexical sequences. The method relied on word n-grams and Jaccard's similarity co-efficient in order to correctly identify the author of extracted email samples. The method was systematically and rigorously tested in an experiment comprising a total of 3,600 tests, attributing samples of 20%, 15%, 10%, 5% and 2% of the twelve EEC12 authors' emails, ranging between an average of 89 and 13,436 tokens in length. Overall, the results of the experiment were promising. Accuracy rates of 92.64% were returned when attributing the 20% samples (between an average of 951 and 13,436 tokens), with bigrams, trigrams and four-grams

achieving 100% accuracy with eleven of the twelve authors tested, with a very large pool of 176 possible authors. The effectiveness of the method did generally decline as the size of the samples being attributed became smaller, and the success rate with 2% samples was only 17.08%. That said, the method did successfully identify Jim Derrick as the author of samples as small as 77, 84 and 109 tokens in length. A subsequent breakdown of the results by author found that the method worked better for some than others, and that it often performed better when applied to authors with smaller dataset sizes, such as Lavorato, Allen and Derrick, than it did for those with far more data such as Germany, Nemec and Steffes. This led to the suggestion that the nature of the author's style may be more important than the size of the sample in terms of the effect on accuracy rates. If forensic linguists accept that authorship attribution is based on a theory of linguistic individuality, then it must also be assumed that different authors' idiolects will be manifest in different ways, and so a measure that might be successful in identifying one author may be unsuccessful in identifying another. However, the fact that the success of the word n-gram approach differs from author to author here highlights that a preoccupation in authorship research with the amount of data being attributed overlooks how well (or not) approaches perform for the individual authors within the datasets being examined. A further breakdown of results by word n-gram length revealed that in terms of number of individual samples attributed, four-grams outperformed the others, with an overall success rate of 70.67%, ranging from attributing 94.16% (113/120) of 20% samples, to 20.83% (25/120) of 2% samples. However, it was found that while the longer n-grams lengths of five and six words performed best with smaller samples, it was shorter n-grams of four, three and even two words which attributed the most 20% samples. In addition to this, it was revealed that certain n-gram lengths performed best for certain authors. While trigrams attributed the most samples for some authors, four-grams, five-grams and six-grams attributed the most for others. These results prompted the argument that while some authors' idiolects may be manifest in longer lexical sequences, others are manifest in shorter ones. It was found that combining the different n-gram lengths produced better results than those achieved using only one measure. The best combination of two measures was found to be four-grams and six-grams, which together had a success rate of 75.67% (454/600). Extending this further, a combination of the four most successful measures (trigrams through to six-grams) successfully identified the authors of 462

of the 600 samples, a success rate of 77%. Following the experiment, Gerald Nemec was used as a case study to examine the specific word n-grams which accounted for the successful attribution of his 5% samples in that they were found in both his disputed sample and the remainder of his 'known' emails. A pool of five-grams was identified which were recurrent in his emails and distinctive of his writing within the Enron corpus. This linguistic evidence was used to support the statistical results. Overall, it was argued that this approach combines the best elements of stylistic and stylometric techniques in attributing authorship. First, word n-grams are a feature which, as suggested by the results of Chapter 5, can be indicative of a person's idiolect, and as such, there is theoretical and linguistic motivation behind their use in identifying authors. Second, the actual attribution of samples was performed through the statistical measurement of similarity between the disputed and known data. As all of the n-grams in the data were used, there was no analyst subjectivity or intuition involved in the comparison of texts. The results of the attributions were expressed in statistical terms, producing known success and error rates. These rates can be considered reliable, having been cross-validated on ten different random samples of each sample size, for each author. Finally, as exemplified by the Nemec case study, this method allows for the statistical results to be explained in linguistic terms, so that the exact nature of stylistic similarity between the disputed and known data is transparent. There are a few methodological caveats for consideration (discussed in Section 6.5). First, it is a priority of future research to identify a reliable threshold of similarity for the Jaccard measurement, such as a specific Jaccard score or level of similarity between texts above which an attribution can be considered reliable and below which any attribution should be considered with caution. Also, the effectiveness of the method is to be assessed when the size of the 'known' comparison documents is restricted. In this thesis only the size of the 'disputed' sample has been controlled, and future research will observe the impact on the method when more or less 'known' data is available for comparison. As these issues are addressed, the method will be continually improved so that it is of optimal accuracy and reliability for forensic application.

Chapter 7 shifted focus from authorship attribution to author profiling. The methodological disadvantages of purely stylometric approaches to the study of authorship revealed by preceding chapters has implications for a field which, to date, has produced only stylometric research. The aim of this chapter was to combine

nomothetic (statistical) and idiographic (stylistic and case study) techniques in analysing groups of authors' use of linguistic features in relation to their social characteristics and identities. The first half of the analysis focused on identifying a set of linguistic features, the relative frequencies of which distinguished between different groups of authors in the EEC80 sample, particularly between male and female employees and those with different occupations within the company. It was found that, of the 291 features used in the analysis, only 12.03% (35/291) showed a statistically significant difference in use across the gender, and only 27.15% (79/291) discriminated between the eight occupation groups. This relatively small number of useful features suggested that the linguistic differences across these groups may not be as great as might be presumed, and so cast into doubt the feasibility of this kind of quantitative author profiling. Furthermore, 17 of the 291 features were found to discriminate between authors both in terms of gender *and* occupation, highlighting the difficulty in determining which particular social characteristic might be considered most 'responsible' for the use of such linguistic features. Following from the nomothetic analysis, three features were selected for closer stylistic study on the basis that they were significant features of a particular gender such as in the case of *for* and *hi* (both female) or both gender and occupation as in the case of *why* (males and traders). These words were used in order to more closely examine the different ways they were used by different groups of authors. In addition, they served as a point of entry for idiographic stylistic comparisons across individual authors who make up these groups. An analysis of these words, as used in collocation by different groups, revealed that the quantitative results offered by the nomothetic procedure were indicative of underlying differences in the linguistic practices of different groups and that, essentially, they tell only part of the story. For example, *for* is used more by males higher up the corporate hierarchy in word clusters in which they thank their recipients for *updates*, compared with females lower down the hierarchy who more frequently use *for* in thanking their recipients for *help*. Similarly, not only do female Enron employees use *hi* more than males, they use it in such a way as to open emails which subsequently include requests or directives. Furthermore, idiographic analyses of how individual authors use these words highlight the risks and inaccuracies in considering social groups as being homogenous. In addition to qualitative differences across the groups, different authors within the *same* group use these features in different ways. Marie Heard and

Kay Mann, for example, are both female lawyers who make frequent use of *hi*. However, while Heard changes the punctuation which she attaches to *hi* greetings in relation to the function of the email and participant she is emailing, Mann does not. Similarly, although Kate Symes and Eric Bass are both traders, Symes uses *why* interrogatives in such a way that relates to her professional workplace identity, while Bass' use of such interrogatives is more closely related to the relationships he has with his audience. Overall, it was argued that authors use particular linguistic features in response to different communicative contexts and functions, and to project different aspects of their identity accordingly, rather than the use of features being determined by whether they are male or female, or CEO or assistant. As such, nomothetic profiling relies on an oversimplified one-dimensional conceptualisation of author identity. In turn, stylistic idiographic analyses can be used to supplement purely statistical findings, and can be used to consider an author's linguistic profile, not just in terms of their social characteristics, but in terms of what kind of communicator they are.

8.3 Contributions and future directions

Because the focus of this thesis was to address the theoretical and methodological challenges in the field of authorship analysis, the contributions it makes are many and varied. Following on from these contributions, a number of suggestions for further research can be proposed.

This study has provided evidence which supports the arguments of forensic linguists that the concepts of Base Rate Knowledge and population-level distinctiveness are ones on which the investigation of idiolectal variation rely. In term of linguistics more generally, the idiolectal evidence offered here adds to the existing corpus linguistic (e.g. Mollin 2009b; Barlow 2010) and sociolinguistic (e.g. Kuhl 2003; Johnstone 1996; 2009) research that has reported results in support of a theory of idiolect. With studies such as this, and as the notion of idiolect develops as a practical and empirically verifiable phenomenon, forensic authorship attribution stands to benefit more than most fields. Further research needs to be undertaken, of course, in evaluating the notions of Base Rate Knowledge and population-level distinctiveness in the investigation of idiolect across different corpora, settings and contexts. The use of corpus linguistics in this way in forensic linguistics relies on the building and sharing of suitable corpora. This study has used the Enron Email

Corpus to represent ‘a relevant population, or group of language users from the same linguistic community’ (Turell and Gavaldà 2013: 499) that can be used to measure the distinctiveness of potentially idiolectal linguistic features. Given its size, the amount of naturally-occurring data available for individual authors, its stability across genre and its free online availability, the Enron Email Corpus may be the best of its kind for forensic purposes. As Kredens and Coulthard (2012: 511–2) note, such useful data for authorship analysis is rare. They give details of specialised corpora of suicide notes (Schneidman and Farbero 1957; Shapero 2011) and SMS text messages (Dyer 2008; Tagg 2009; Grant 2010), which may be used for forensic purposes. However, it is not known whether these corpora are freely available. Another useful corpus is the ‘Blog Authorship Corpus’, comprising over 19,000 bloggers and 140 million words, compiled and distributed by Schler et al. (2006). Such large corpora of online texts are relatively easy to collect now with data mining and programming techniques. Jack Grieve and his colleagues at Aston University are in the process of collecting a huge corpus of billions of words of tweets, for example (*The Telegraph* 2014). While this is not being built specifically for forensic purposes, should the corpus be organised in such a way that tweets of individual people can be identified, this would offer a fantastic dataset for authorship analysis of short form messages. In the corpus linguistic community, the importance of retaining and coding of metadata is being increasingly stressed (e.g. Murphy and Knight 2014). It can be hoped that in the development of new corpora in corpus linguistics, the language use of individual writers or speakers is demarcated in such a way as to be useful for authorship analysts. One particularly exciting development in this regard is the creation of the *Spoken British National Corpus 2014*, by Lancaster and Cambridge Universities (*CASS* 2014). Although spoken rather than written, recorded submissions from individual speakers and participants may create a useful dataset for the analysis of idiolect.

In particular in this study, the idiolectal nature of collocations has been observed. On the one hand, this provides evidence in favour of those who argue that lexical co-selections are unique to individual language users (Nattinger and DeCarrico 1992; Wray 2002; Hoey 2005). On the other, it offers collocations and word n-grams to forensic authorship attribution research and practice as features of linguistic variation that can be exploited in distinguishing between authors and attributing disputed documents. Word n-grams have been used (under various

guises) in authorship analysis (e.g. Hoover 2002, 2003; Coyotl-Morales et al. 2006; Sanderson and Guenter 2006; Grieve 2007; Juola 2013; Larner 2014). However, these studies have not tested so many different lengths of n-gram, or on so many authors or on different sample sizes. Given the suggested idiolectal nature of collocations, they clearly offer potential for further scrutiny by authorship analysts, over a range of corpora. They perhaps offer a more progressive future than persisting with relative frequencies of function words which have dominated the landscape over the last few decades.

In terms of author profiling, it is hoped that the results of this study offer points for caution in the stylometric correlation of linguistic features and discrete, one-dimensional social categories. The analyses here have introduced idiographic or stylistic approaches into an author profiling context for the first time, demonstrating the value they can add to critically examining nomothetic results. More theoretically, though, the main contribution to this particular field is potentially the emphasis on the multi-dimensional multi-faceted nature of author identity. The argument that 'identity is something we actively do, rather than something we passively are' (Angouri and Marra 2011: 1), and that social factors such as gender, age and ethnicity 'are as resources that speakers use to create unique voices, than determinants of how they will talk' (Johnstone 1996: 11) are well-established in sociolinguistics. Yet, such nuanced relationships between aspects of identity and language use are critically overlooked by quantitative approaches to profiling. Thankfully, this conceptualisation of the linguistic individual is making its way into forensic linguistics. As Kredens (personal communication 2014) remarks:

> viewing speakers as active agents drawing on the resources they have at their disposal is a more promising (but also much more challenging) approach in our discipline, not least because it is more conducive to addressing the various issues to do with the validity of findings.

Finally, the most important contribution of this work in terms of advancing the science of authorship analysis is that it actively seeks to combine theoretical and methodological aspects of stylometric and stylistic approaches. In doing so, it adds to very recent work which also makes such advances (Grant 2010; 2013; Nini and Grant 2013; Argamon and Koppel 2010; 2013). Such work is vital within the current climate of authorship research, lest the field continue to produce divergent methodologies for years to come. In October 2012, Lawrence Solan, Professor of

Law at Brooklyn Law School, hosted an Authorship Attribution Workshop bringing together experts from stylistic and stylometric standpoints. The resulting special issue of *The Journal of Law and Policy* (Volume 21, Issue 2) published the most cutting edge research from both approaches, as well as comments from legal experts, many of the papers from which have been cited throughout this thesis. Events of this kind are invaluable in starting conversations across disciplines that will potentially benefit the field so immeasurably. In his summarising article entitled *Intuition* versus *Algorithm*, (the name of which motivated the title of this thesis), Solan (2013: 576) remarks: 'I firmly believe that far more collaboration among scholars with different areas of expertise is absolutely essential'. This sentiment is echoed here, along with a need to provide (forensic) linguists with formal training in statistics and rudimentary computational linguistics.

8.4 Closing remarks

In the constitution of the *International Association of Forensic Linguists*, one of the purposes of the association is listed as being 'research into the practice, improvement, and ethics of expert testimony and the presentation of linguistic evidence' (IAFL 2013). This study has endeavoured to pursue this aim. It has offered corpus linguistics as a means through which theoretical concepts can be empirically tested, and divergent techniques can be combined, in the analysis of forensic texts. This thesis opened with a quotation from Svartik (1968: preface) which stated that forensic linguistics offers the linguist a rare opportunity 'of making a contribution that might be directly useful to society'. If a forensic linguist's contributions to the domains of law enforcement and security settings are to truly benefit society, then they must be of the highest possible scientific standards. It is hoped that this study represents at least a small step in this direction.

References

Abbasi, Ahmed and Hsinchun Chen. 2005. Applying authorship analysis to extremist-group web forum messages. *IEEE Intelligent Systems* 20(5), 67–75.

Aijmer, Karin. 1996. *Conversational Routines in English: Convention and Creativity*. London: Longman.

Aijmer, Karin. 1997. *I think* — an English modal particle. In Toril Swan and Olaf Jansen Westvik (eds.) *Modality in Germanic languages. Historical and comparative perspectives*. Berlin: Mouton, 1–47.

Al Sallab, Ahmad A. and Mohsen A. A. Rashwan. 2012. E-mail classification using deep networks. *Journal of Theoretical and Applied Information Technology* 37(2), 241–251.

Alison, Laurence. 2005. From trait-based profiling to psychological contributions to apprehension methods. In Laurence Alison (ed.) *The Forensic Psychologists' Casebook: Psychological Profiling and Criminal Investigation*. London: Routledge, 3–22.

Alison, Laurence, Craig Bennell, Andreas Mokros and David Omerod. 2002. The personality paradox in offender profiling. A theoretical review of the processes involved in deriving background characteristics from crime scene actions. *Psychology, Public Policy, and Law* 8(1), 115–135.

Angouri, Jo and Meredith Marra. 2011. *Constructing Identities at Work.* London: Palgrave.

Anthony, Laurence. 2014. *AntConc* (Version 3.4.1). Tokyo, Japan: Waseda University. Available from http://www.antlab.sci.waseda.ac.jp [Accessed January 2014].

Archer, Dawn. (ed.). 2009. *What's in a Word-list? Investigating Word Frequency and Keyword Extraction*. Farnham: Ashgate.

Archer, Dawn, Jonathan Culpeper and Paul Rayson. 2009. Love – 'a familiar or a devil'? An exploration of key domains in Shakespeare's comedies and tragedies. In Dawn Archer (ed.) *What's in a Word-list? Investigating Word Frequency and Keyword Extraction*. Farnham: Ashgate, 137–158.

Argamon, Shlomo, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. 2003. Gender, genre, and writing style in formal written texts. *Text* 23(3), 321–346.

Argamon, Shlomo and Shlomo Levitan. 2005. Measuring the usefulness of function words for authorship attribution. In *Proceedings of ACH/ALLC Conference*, University of Victoria, BC, Association for Computing and the Humanities, 1–3.

Argamon, Shlomo and Sushant Dhawle, Moshe Koppel and James W. Pennebaker. 2005. Lexical predictors of personality type. In *Proceedings of the 2005 Classification Society of North America Annual Meeting*. [online]. Available from: http://www.lingcog.iit.edu/wp-content/papercite-data/pdf/argamon-etal-csna.pdf. [Accessed February 2013].

Argamon, Shlomo, Moshe Koppel, James W. Pennebaker, Jonathan Schler. 2009. Automatically profiling the author of an anonymous text. *Communication of the ACM* 52(2), 119–123.

Argamon, Shlomo and Moshe Koppel. 2010. The rest of the story: Finding meaning in stylistic variation. In Shlomo Argamon, Kevin Burns and Shlomo Dubnov (eds.) *The Structure of Style. Algorithmic Approaches to Understanding Manner and Meaning*. London: Springer, 79–112.

Argamon, Shlomo and Moshe Koppel. 2013. A systemic functional approach to automated authorship analysis. *Journal of Law and Policy* 21(2), 299–316.

Austin, John, L. 1962. *How to do Things with Words*. Oxford: Oxford University Press.

Baayen Harald, Hans van Halteren and Fiona Tweedie. 1996. Outside the cave of shadows: using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing* 11(3), 121–132.

Baker, Paul. 2010. *Sociolinguistics and Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Baker, Paul. 2014. *Using Corpora to Analyze Gender*. London: Bloomsbury.

Baker, Paul, Costas Gabrielatos, Majid Khosravinik, Michał Krzyzanowski, Tony McEnery and Ruth Wodak. 2008. A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse and Society* 19(3), 273–306.

Baker, Paul, Costas Gabrielatos and Tony McEnery. 2013. Sketching Muslims: a corpus driven analysis of representations around the word 'Muslim' in the British press 1998–2009. *Applied Linguistics* 34(3), 255–278.

Bamman, David, Jacob Eisenstein and Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics* 18(2), 135–160.

Barbieri, Federica. 2008. Patterns of age-based linguistic variation in American English. *Journal of Sociolinguistics* 12(1), 58–88.

Barlow, Michael. 2010. Individual usage: a corpus-based study of idiolects. Paper presented at the 34th International LAUD Symposium, Landau, Germany. [online]. Available from http://michaelbarlow.com/barlowLAUD.pdf [Accessed December 2011].

Barlow, Michael. 2013. Individual differences and usage-based grammar. *International Journal of Corpus Linguistics* 18(4), 443–478.

Barnbrook, Geoff, Oliver Mason and Ramesh Krishnamurthy. 2013. *Collocation: Applications and Implications*. London: Palgrave.

Bartol Curt, R. and Anne M. Bartol. 2008. *Introduction for Forensic Psychology: Research and Application* (2nd edn). London: Sage.

Baucus, Max and Charles E. Grassley. 2003. *Report of Investigation of Enron Corporation and Related Entities Regarding Federal Tax and Compensation Issues, and Policy Recommendations. Volume III: Appendices C & D*. [online]. Available from: http://www.gpo.gov/fdsys/pkg/GPO-CPRT-JCS-3-03/content-detail.html. [Accessed May 2013].

Bax, Ingrid, Pufahl. 1986. How to assign work in an office. A comparison of spoken and written directives in American English. *Journal of Pragmatics10*, 673–692.

Baxter, Judith. 2012. Women of the corporation: A sociolinguistic perspective of senior women's leadership language in the UK. *Journal of Sociolinguistics* 16(1), 81–107.

BBC News. 2002. Enron Scandal at-a-glance. 22 August 2002. [online]. Available from http://news.bbc.co.uk/1/hi/business/1780075.stm. [Accessed February 2012].

BBC News. 2008. The case for forensic linguistics. 8 September 2008. [online]. Available from: http://news.bbc.co.uk/1/hi/sci/tech/7600769.stm]. [Accessed 16 August 2014].

BBC, World Service, The. 2011. When to use ‘will’, ‘shall’, ‘would’ and ‘should’. [online]. Available from http://www.bbc.co.uk/worldservice/learningenglish/grammar/learnit/learnitv43.shtml. [Accessed July 2014].

Becker, Joseph, D. 1975. The phrasal lexicon. In Bonnie, L. Nash-Webber and Roger Shank. *Theoretical Issues in Natural Language Processing*. Cambridge, MA: Bolt Beranek and Newman, 60–63. [online]. Available from: http://ww2.cs.mu.oz.au/acl/T/T75/T75-2013.pdf. [Accessed January 2014].

Bekkerman, Ron, Andrew McCallum and Gary Huang. 2004. *Automatic Categorization of Email into Folders: Benchmark Experiments on Enron and SRI Corpora.* Technical Report for the Center of Intelligent Information Retrieval. University of Massachusetts – Amherst. [online] Available from http://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1217&context=cs_faculty_pubs. [Accessed April 2012].

Bennell, Craig and Natalie J. Jones. 2005. Between a ROC and a hard place: a method for linking serial burglaries by *modus operandi. Journal of Investigative Psychology and Offender Profiling* 2(1), 23–41

Bennell, Craig, and Jessica Woodhams. 2012. Behavioural linking of crimes. *Journal of Investigative Psychology and Offender Profiling* 9(3), 199–200.

Bennell, Craig, Rebecca Mugford, Holly Ellingwood and Jessica Woodhams. 2014. Linking crimes using behavioural clues: Current levels of linking accuracy and strategies for moving forward. *Journal of Investigative Psychology and Offender Profiling* 11(1), 29–56.

Benston, George, J. and Al L. Hartgraves. 2002. Enron: What happened and what we can learn from it. *Journal of Accounting and Public Policy* 21(2), 105–127.

Berba-Sardinha, Tony. 2000. Comparing corpora with WordSmith tools: how large must the reference corpus be? In *Proceedings of the Workshop on Comparing Corpora* 9, 7–13.

Berman, Dennis K. 2003. Government posts Enron's e-mail. *Wall Street Journal.* 6 October 2003. [online]. Available from: http://online.wsj.com/articles/SB106540100255454500. [Accessed March 2013].

Biber, Douglas. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, Douglas and Edward Finegan. 1994. Introduction: Situating register in sociolinguistics. In Douglas Biber and Edward Finegan (eds.) *Sociolinguistic Perspectives on Register*. Oxford: Oxford University Press, 3–12.

Biber, Douglas and Susan Conrad. 2009. *Register, Genre and Style.* Cambridge: Cambridge University Press.

Binongo, José, Nilo G. and M.W.A Smith. 1999. The application of principle component analysis to stylometry. *Literary and Linguistic Computing* 14(4), 445–466.

Bjørge, Anne Kari. 2007. Power distance in English lingua franca email communication. *International Journal of Applied Linguistics* 17(1), 60–80.

Bloch, Bernard. 1948. A set of postulates for phonemic analysis. *Language* 24(1), 3–46.

Bloomfield, Leonard. 1933. *Language*. New York: Holt, Rinehart, and Winston.

Bou-Franch, Patricia. 2006. Solidarity and deference in computer-mediated communication: A discourse-pragmatic analysis of students' emails to lecturers. In Patricia Bou-Franch (ed.) *Ways into Discourse*. Granada: Comares, 74–94.

Bou-Franch, Patricia. 2011. Openings and closings in Spanish email conversations. *Journal of Pragmatics* 43(6), 1772–1785.

Boulis, Constantinos and Maro Ostendorf. 2005. A quantitative analysis of lexical differences between genders in telephone conversations. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, 435–442. [online]. Available from: http://www-ssli.ee.washington.edu/people/boulis/gender_ACL05.pdf [Accessed March 2014].

Bremner, Julie, Christopher L.G. Frid, and Stuart I. Rogers. 2003. Assessing marine ecosystem health: The long-term effects of fishing on functional biodiversity in North Sea benthos. *Aquatic Ecosystem Health and Management* 6(2), 131–137.

Brezina, Vaclav and Miriam Meyerhoff. 2014. Significant or random?: A critical review of sociolinguistic generalisations based on large corpora. *International Journal of Corpus Linguistics* 19(1), 1–28.

Brocardo, Marcelo Luiz, Issa Traore, Sherif Saad, and Isaac Woungang. 2013. Authorship verification for short messages using stylometry. In *Proceedings of the IEEE International Conference on Computer, Information and Telecommunication Systems.*[online]. Available from: http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=6705711&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D6705711. [Accessed March 2014].

Brown, Penelope and Stephen C. Levinson. 1987. *Politeness: Some Universals in Language Usage*. Cambridge: Cambridge University Press.

Brysbaert, Marc, Boris New and Emmanuel Keuleers. 2012. Adding part-of-speech information to the SUBTLEX-US word frequencies. *Behavioural Research Methods* 44(4), 991–997.

Burger, John D., John Henderson, George Kim and Guido Zarrella. 2011. Discriminating gender on Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 1301–1309. [online]. Available from: http://www.aclweb.org/anthology/D11-1120. [Accessed March 2014].

Burrows, John. 2002. 'Delta:' A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing* 17(3), 267–287.

Burrows, John. 2003. Questions of authorship: attribution and beyond. *Computers and the Humanities* 37(1), 1–26.

Burrows, John. 2005. Andrew Marvell and the 'Painter Satires': A computational approach to their authorship. *Modern Language Review* 100(2), 281–97.

Burrows, John. 2007. All the way through: Testing for authorship in different frequency strata. *Literary and Linguistic Computing* (22)1, 27–47.

Business Week Magazine. 2001. The Fall of Enron. 16 December 2001. [online]. Available from: http://www.businessweek.com/stories/2001-12-16/the-fall-of-enron [Accessed April 2013].

Butters, Ron. 2012. Retiring President's closing address: ethics, best practices, and standards. In Samuel Tomblin, Nicci MacLeod, Rui Sousa-Silva and Malcolm Coulthard (eds.) *Proceedings of the Tenth International Association of Forensic Linguists' Biennial Conference*, Aston University, Birmingham, 351–361. [online]. Available from: www.forensiclinguistics.net [Accessed November 2012].

Bybee, Joan. 2006. From usage to grammar: The mind's response to repetition. *Language* 82(4), 711–733.

Cameron, Deborah. 2008. *The Myth of Mars and Venus: Do men and Women Really Speak Different Languages?* Oxford: Oxford University Press.

Can, Falzi and Jon M. Patton. 2004. Change of writing style with time. *Computers and the Humanities* 38(1), 61–82.

Canter, David. 1995. The psychology of offender profiling. In David Carson and Ray Bull (eds.) *Handbook of Psychology in Legal Contexts* (2nd edn). Chichester: Wiley, 343–355.

Canter, David and Rupert Heritage. 1990. A multivariate model of sexual offence behaviour: Developments in 'offender profiling'. I. *The Journal of Forensic Psychiatry* 1(2), 18 –212.

Canter, David, and Laurence Alison. 2000. Profiling property crimes. In David Canter and Laurence Alison (eds.) *Profiling Property Crimes*. Burlington, VT: Ashgate, 1–30.

Canter, David, Craig Bennell, Laurence Alison and Steve Reddy. 2003. Differentiating sex offences: A behaviorally based thematic classification of stranger rapes. *Behavioural Sciences and the Law* 21(2), 157–174.

Canter, David and Donna Youngs. 2009. *Investigative Psychology: Offender Profiling and the Analysis of Criminal Action*. Chichester: Wiley.

CASS (*ESRC Centre for Corpus Approaches to Social Science, Lancaster University*). 2014. Spoken BNC2014 project announcement. [online]. Available from: http://cass.lancs.ac.uk/?p=1335. [Accessed August 2014].

Champod, Cristophe and Ian W. Evett. 2000. Comentary on Broeders. *Forensic Linguistics. The International Journal of Speech Language and the Law* 7(2), 238–243.

Chaski, Carole, E. 1997. Who wrote it? Steps toward a science of authorship identification.' *National Institute of Justice Journal* 233, 15–22.

Chaski, Carole, E. 2001 Empirical evaluations of language-based author identification techniques. *Forensic Linguistics: (The International Journal of Speech Language and the Law)* 8(1), 1–65.

Chaski, Carole, E. 2005. Who's at the keyboard? Authorship attribution in digital evidence investigations. *International Journal of Digital Evidence* 4(1), 1–14.

Chaski, Carole, E. 2013. Best practices and admissibility of forensic author identification. *Journal of Law and Policy* 21(2), 333–376.

Chen, Xiaoling, Peng Hao, Rajarathnam Chandramouli and K. P. Subbalakshmi. 2011. Authorship Similarity Detection from Email Messages. In Petra Perner (ed.) *Machine Learning and Data Mining (MLDM) in Pattern Recognition*, 375–386.

Cheng, Edward, K. 2013. Being pragmatic about forensic linguistics. *Journal of Law and Policy* 21(2), 541–550.

Clapham, Christopher and James Nicholson. 2009. *The Concise Oxford Dictionary of Mathematics*. Oxford: Oxford University Press.

Clark, Alexander, M.S. 2011. Forensic stylometric authorship analysis under the Daubert Standard. *Law and Literature eJournal*. [online]. Available from: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2039824. [Accessed April 2012].

Coates, Jennifer. 2004. *Women, Men and Language: A Sociolinguistic Account of Gender Differences in Language* (3rd edn.). London: Longman.

Cohen, William W. 2009. *Enron Email Dataset*. [online]. Available from: http://www.cs.cmu.edu/~enron/. [Accessed November 2010].

Cohn, Michael, A., Matthias R. Mehl, and James W. Pennebaker. 2004. Linguistic markers of psychological change surrounding September 11, 2001. *Psychological Science* 15(10), 687–693.

Conley, John, M. and William M. O'Barr. 1998. *Just Words: Law, Language and Power*. Chicago: University of Chicago Press.

Cotterill, Janet. 2010. How to use corpus linguistics in forensic linguistics. In Anne O'Keefe and Michael McCarthy (eds.) *The Routledge Handbook of Corpus Linguistics.* London: Routledge, 578–590.

Coulmas, Florian. 1979. On the sociolinguistic relevance of routine formulae. *Journal of Pragmatics* 3, 239–266.

Coulmas, Florian. 1981. Introduction: conversational routine. In Florian Coulmas (ed.) *Conversational Routine: Explorations in standardized communication situations and prepatterned speech*. The Hague: Mouton, 1–17.

Coulthard, Malcolm. 1994. On the use of corpora in the analysis of forensic texts. *Forensic Linguistics. The International Journal of Speech, Language and the Law* 1(1), 27–43.

Coulthard, Malcolm. 2004. Author identification, idiolect, and linguistic uniqueness. *Applied Linguistics* 24(4), 431–447.

Coulthard, Malcolm. 2010. Experts and opinions: In my opinion. In Malcolm Coulthard and Alison Johnson (eds.) *The Routledge Handbook of Forensic Linguistics*. London: Routledge, 473–486.

Coulthard, Malcolm. 2013. On admissible linguistic evidence. *Journal of Law and Policy* 21(2), 441–466.

Coulthard, Malcolm and Alison Johnson. 2007. *An Introduction to Forensic Linguistics: Language in Evidence*. London: Routledge.

Coulthard, Malcolm, Tim Grant, and Krzysztof Kredens. 2011. Forensic Linguistics. In Ruth Wodak, Barbara Johnstone and Paul Kerswill (eds.) *The SAGE Handbook of Sociolinguistics*. London: Sage, 531–544.

Coyotl-Morales, Rosa. M., Luis Villaseñor-Pineda, Manuel Montes-y-Gómez and Paolo Rosso. 2006. Authorship attribution using word sequences. In *Proceedings of the 11th Iberoamerican Congress on Pattern Recognition*. Berlin: Springer, 844–853.

Creamer, Germán, Ryan Rowe, Shlomo Hershkop and Salvatore, J, Stolfo. 2009. Segmentation and automated social hierarchy detection through email network analysis. In Haizheng Zhang, Myra Spiliopoulou, Bamshad Mobasher, C. Lee Giles, Andrew McCallum, Olfa Nasraoui, Jaideep Srivastava, John Yen (eds.) *Advances in Web Mining and Web Usage Analysis*, Berlin: Springer, 40–58.

Crystal, David. 1986. I shall and I will. *English Today* 2(1), 42–66.

Culpeper, Jonathan. 2009. Keyness: words, parts-of-speech and semantic categories in the character-talk of Shakespeare's Romeo and Juliet. *International Journal of Corpus Linguistics* 14(1), 29–59.

Culwin, Fintan and Mike Child. 2010. Optimising and automating the choice of search strings when investigating possible plagiarism. In *Proceedings of the 4th International Plagiarism Conference.* [online] Available from: http://www.plagiarismadvice.org/researchpapers/ item/optimising-and-automating-the-choice-of-search-strings-wheninvestigating-possible-plagiarism [Accessed July 2013].

Davies, Mark. 2012. *The Corpus of Contemporary American English: 450 million words,1990-presen*t [online]. http://corpus.byu.edu/coca/. [Accessed December 2011].

Davies, Mark and Dee Gardner. 2010. *A Frequency Dictionary of Contemporary American English: Word Sketches, Collocates and Thematic Lists*. London: Routledge.

de Vel, Olivier, Alison Anderson, Malcolm Corney, and George Mohay. 2001. Mining e-mail content for author identification forensics. *Association for Computing Machinery Sigmod Record* 30(4), 55–64.

Decamp, David. 1969. *Toward a Formal Theory of Sociolinguistics*. Unpublished Manuscript. University of Texas, Austin.

Deng, Fan, Stefan Siersdorfer, Sergej Zerr. 2012. Efficient Jaccard-based Diversity Analysis of Large Document Collections. *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM '12),* 1402–1411.

Diederich, Joachim, Jörg Kindermann, Edda Leopold, Gerhard Paass. 2003. Authorship attribution with Support Vector Machines. *Applied Intelligence* 19(1/2), 109–123.

Dikmen, Mert, and Thomas Huang. 2011. Leveraging social network information to recognize people. Paper presented at *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops.* Colorado Springs, USA, June 2011.

Dittmar, Norbert. 1996. Explorations in 'Idiolects'. In Robin Sackmann and Monika Budde (eds). *Theoretical Linguistics and Grammatical Description: Papers in honour of Hans-Heinrich Lieb*. Amsterdam: John Benjamins, 109–128.

Dörnyei, Zoltán. 2007. *Research Methods in Applied Linguistics: Quantitative, Qualitative and Mixed Methodologies*. Oxford: Oxford University Press.

Downing, Bruce, T. 1969. Vocatives and third-person imperatives in English. *Papers in Linguistics* 1(3), 570–592.

Drew, Paul and John Heritage. 1992. Analysing talk at work: an introduction. In Paul Drew and John Heritage (eds.) *Talk at Work: Interaction in institutional settings*. Cambridge: Cambridge University Press, 3–65.

Dunning, Ted. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19(1), 61–74.

Dyer, Wynde. 2008. Just text me: The preliminary results of a text message corpus compilation pilot project. Paper presented at *the Fourth International Conference on Technology, Knowledge and Society*. Boston, USA, January 2008.

Eckert, Penelope. 1989. The whole woman: Sex gender differences in variation. *Language Variation and Change* 1(1), 245–267.

Eckert, Penelope. 2008. Variation in the indexical field. *Journal of Sociolinguistics* 12(4), 453–476.

Eckert, Penelope and Sally McConnell-Ginet. 1992. Thing practically and look locally: Language and gender as community-based practice. *Annual Review of Anthropology* 21, 461–490.

Eckert, Penelope and Sally McConnell-Ginet. 1998. Communities of practice: where language, gender and power all live? In Jennifer Coates (ed.) *Language and Gender: A Reader*. Oxford, Blackwell, 484–494.

Eckert, Penelope and Sally McConnell-Ginet. 2003. *Language and Gender*. Cambridge: Cambridge University Press.

Ecology. 2013. Volume 95, Issue 3. Forum on '*p* values and model selection'. (March 2013).

Economidou-Kogetsidis Maria. 2011. 'Please answer me as soon as possible': Pragmatic failure in non-native speakers' e-mail requests to faculty. *Journal of Pragmatics* 43, 3193–3215.

Economist, The. Shall we try that again? 20 May 2011. [online]. Available from: http://www.economist.com/blogs/johnson/2011/05/modal_verbs [Accessed July 2014].

Eder, Maciej. 2013. Does size matter? Authorship attribution, small samples, big problem. *Literary and Linguistic Computing* [online]. Available from: http://llc.oxfordjournals.org/content/early/2013/11/14/llc.fqt066.full [Accessed June 2014].

EDRM. The Electronic Discovery Reference Model. 2013. *EDRM Enron Email Data Set v2*. [online]. Available from: http://www.edrm.net/resources/data-sets/edrm-enron-email-data-set-v2. [Accessed April 2013].

Ehrlich, Susan and Alice, F. Freed. 2010. The function of questions in institutional discourse: an introduction. In Susan Ehrlich and Alice F. Freed (eds.) *Why Do You Ask?: The Function of Questions in Institutional Discourse*. Oxford: Oxford University Press, 3–19.

Ellegard, Alvar, A. 1962. *Statistical Method for Determining Authorship: The Junius Letters, 1769–1772*. Gothenburg: University of Gothenburg.

Estival, Dominique, Tanja Gaustad, Son Bao Pham, Will Radford and Ben Hutchinson. 2007. Author profiling for English emails. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*, 263–272. [online]. Available from: http://aclweb.org/anthology/D/D11/D11-1148.pdf [Accessed February 2014].

Evans, Melanie. 2013. *The Language of Queen Elizabeth I: A sociolinguistic Perspective on Royal Style and Identity*. Oxford: Wiley-Blackwell.

Evett, Ian, W. 1993 Establishing the evidential value of a small quantity of material found at a crime scene. *Journal of the Forensic Science Society* 33(2), 83–86.

Federal Energy Regulatory Commission (FERC). 2013. *Information Released in Enron Investigation*. [online]. Available from http://www.ferc.gov/industries/electric/indus-act/wec/enron/info-release.asp#skipnav. [Accessed April 2013].

Fetzer, Anita. 2008. 'And i think that is a very straightforward way of dealing with it'. The communicative function of cognitive verbs in political discourse. *Journal of Language and Social Psychology* 27(4), 384–396.

Fetzer, Anita. 2014. *I think, I mean* and *I believe* in political discourse. *Functions of Language* 21(1), 67–94.

Financial Times, The. 2002. Enron – The Background. Timeline: Enron's rise and fall. [online]. Available from: http://specials.ft.com/enron/FT38O0NQUWC.html. [Accessed February 2012].

Fiore, Andrew and Jeff Heer. 2013. *UC Berkeley Enron Email Analysis*. [online]. Available from: http://bailando.sims.berkeley.edu/enron_email.html. [Accessed April 2013].

Firth, John Rupert. 1957. A synopsis of linguistic theory 1930–1955. In Frank Robert Palmer (ed.*) Selected papers of J.R. Firth 1952–1959*. London: Longman, 168–205.

Fitzgerald, Jim. R. 2004. Using a forensic linguistic approach to track the Unabomber. In John H. Campbell and Don Denivi (eds.) *Profilers*: *Leading Investigators take you Inside the Criminal Mind*. New York: Prometheus Books, 193–222.

Flowerdew, Lynne. 2004. The argument for using English specialized corpora to understand academic and professional settings. In Ulla Connor and Thomas A. Upton (eds.) *Discourse in the Professions: Perspectives from Corpus Linguistics*. Amsterdam: John Benjamins, 11–33.

Foulkes, Paul and Jennifer Hay. 2013. The evolution of medial /t/ over real and imagined time. Paper presented at *UK Language Variation and Change 9 (UKLVC9*), Sheffield, England, September 2013.

Garside, Roger and Nicholas Smith. 1997 A hybrid grammatical tagger: CLAWS4. In Roger Garside, Geoffrey Leech and Tony McEnery (eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London: London, 102–121.

Geng, Liqiang, Larry Korba, Xin Wang, Yunli Wang, Hongyu Liu and Yonghua You. 2008. Using data mining methods to predict personally identifiable information in emails. Paper presented at *4th International Conference on Advanced Data Mining and Applications*. Chengdu, China, October 2008. [online]. Available from: http://www.ucalgary.ca/wangx/files/wangx/PII-detection-Enron-ADMA-final.pdf. [Accessed December 2013].

Gibbons, John. 2003. *Forensic Linguistics: An Introduction to Language in the Justice System*. London: Blackwell.

Gilbert, Eric. 2012. Phrases that signal workplace hierarchy. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. [online]. Available from: http://comp.social.gatech.edu/papers/cscw12.hierarchy.gilbert.pdf. [Accessed June 2013].

Gilquin, Gaëtanelle. 2010. *Corpus, Cognition and Causative Constructions*. Amsterdam: John Benjamins.

Gliwa, Bogdan, Anna Zygmunt and Aleksander Byrski. 2012. Graphical analysis of social group dynamics. In *Proceedings of the 2012 4th International Conference on Computational Aspects of Social Networks, (CASoN).* [online]. Available from: http://arxiv.org/pdf/1303.6088v2.pdf. [Accessed March 2013].

Grant, Tim. 2004. *Authorship Attribution in a Forensic Context*. PhD thesis. University of Birmingham. [online]. Available from: http://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.529439. [Accessed October 2011].

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

Grant, Tim. 2008. Approaching questions in forensic authorship analysis. In John Gibbons and M. Teresa Turell (eds.) *Dimensions of Forensic Linguistics*. Amsterdam: John Benjamins, 215–229.

Grant, Tim. 2010. Txt 4n6: Idiolect free authorship analysis? In Malcolm Coulthard and Alison Johnson (eds.) *The Routledge Handbook of Forensic Linguistics.* London: Routledge, 508–522.

Grant, Tim. 2013. Txt 4N6: Method, consistency and distinctiveness in the analysis of SMS text messages. *Journal of Law and Policy* 21(2), 467–494.

Grant, Tim and Kevin Baker. 2001. Identifying reliable, valid markers of authorship: a response to Chaski. *Forensic Linguistics: (The International Journal of Speech Language and the Law)* 8(1), 66–79.

Gries, Stefan Th. 2009. *Statistics for Linguistics with R: A Practical Introduction*. Berlin: Mouton de Gruyter.

Gries, Stefan, Th. 2013. 50-something years of work on collocations: What is or should be next... *International Journal of Corpus Linguistics* 18(1), 137–165.

Grieve, Jack. 2007. Quantitative authorship attribution: An evaluation of techniques. *Literary and Linguistic Computing* 22(3), 251–270.

Grieve, Jack. 2013. Comparative authorship analysis in the Starbucks case. Paper presented at *Current Trends in Forensic Linguistics* conference, Aston University, Birmingham, November 2013.

Häkkänen, Helinä, Petri Lindlöf and Pekka Santtila. Crime scene actions and offender characteristics in a sample of Finnish stranger rapes. *Journal of Investigative Psychology and Offender Profiling* 1(1), 17–32.

Hall, Kathleen Currie. and Amanda Boomershine. 2006. Life, the critical period: An exemplar-based model of language learning. Paper presented at 10th Conference on Laboratory Phonology, Paris, France, June 2006. [online]. Available from: http://www.ling.ohio-state.edu/~kchall/KCH_Third_Year_Paper_revised.pdf [Accessed January 2014].

Halliday, M.A.K. 1989. *Language, Context and Text. Aspects of Language in a Social-semiotic Perspective.* Oxford: Oxford University Press. (As cited by Turell 2010).

Halliday, M.A.K. and Ruqaiya Hasan. 1985. *Language, Context, and Text: Aspects of Language in a Social-Semiotic Perspective.* Oxford: Oxford University Press.

Halliday, M.A.K. and Christian Matthiessen. 2004. *An Introduction to Functional Grammar* (3rd edn.). London: Routledge.

Handford Michael and Petr Matous. 2011. Lexicogrammar in the international construction industry: A corpus-based case study of Japanese-Hong-Kongese on-site interactions in English. *English for Specific Purposes* 30(2), 87–100.

Harris, Sandra. 1984. Questions as a mode of control in magistrates' courts. *International Journal of the Sociology of Language* 49, 5–27.

Hasan, Ruqaiya. 1985. The structure of a text. In M.A.K Halliday and Ruqaiya Hasan (eds.) *Language, Context and Text: Aspect of Language in a Social-Semiotic Perspective*. Cambridge: Cambridge University Press, 52–69.

Hay, Jennifer and Joan Bresnan. 2006. Spoken syntax: The phonetics of "giving a hand" in New Zealand English. *The Linguistic Review* 23, 321–349.

Healy, Paul, M. and Krishna G. Palepu. 2003. The Fall of Enron. *Journal of Economic Perspectives* 17(2), 3–26.

Heritage, John. 2004. Conversation analysis and institutional talk: analysing data. In David Silverman (ed.) *Qualitative Research: Theory, Method and Practice*. London: Sage, 222–243.

Herring, Susan, C. and John C. Paolillo. 2006. Gender and genre variation in weblogs. *Journal of Sociolinguistics* 10(4), 439–459.

Heylighen, Francis and Jean-Marc Dewaele. 2002. Variation in the contextuality of language: an empirical measure. *Foundations of Science* 7(3), 292–340.

Hirst, Graeme and Ol'ga Feiguina. 2007. Bigrams of syntactic labels for authorship discrimination of short texts. *Literary and Linguistic Computing* 22(4), 405–417.

Ho-Abdullah, Imran. 2010. *Variety and Variability: A Corpus-based Cognitive Lexical-semantics Analysis of Prepositional Usage in British, New Zealand and Malaysian English*. Bern: Peter Lang.

Hockett, Charles, F. 1958. *A Course in Modern Linguistics*. New York: The Macmillan Company.

Hockey, Susan. 1988. *Micro Oxford Concordance Program (OCP version 2*). Oxford: Oxford University Press.

Hoey, Michael. 2005. *Lexical Priming: A new theory of words and language*. London: Routledge.

Hoey, Michael. 2006. Language as choice: what is chosen? In Geoff Thompson and Susan Hunston (eds.) *System and Corpus: Exploring Connections*. London: Equinox, 37–53.

Holmes, David, I. 1994. Authorship attribution. *Computers and the Humanities* 28(2), 87– 106.

Holmes, David, I. and Richard S. Forsyth. 1995. The Federalist revisited: New directions in authorship attribution. *Literary and Linguistic Computing* 10(2), 111–127.

Holmes, David, I., Lesley J. Gordon and Christine Wilson. 2001. A widow and her soldier: Stylometry and the American Civil War. *Literary and Linguistic Computing* 16(4), 403–420.

Holmes, Janet. 2000. Doing collegiality and keeping control at work: small talk in government departments. In Justine Coupland (ed.) *Small Talk.* London: Routledge, 32–61.

Holmes, Janet. 2001. *An Introduction to Sociolinguistics* (2nd edn.). London: Longman.

Holmes, Janet. 2006. *Gendered Talk at Work: Constructing Social Identity Through Workplace Interaction*. London: Blackwell.

Holmes, Janet and Meredith Marra. 2002. Having a laugh at work: how humour contributes to workplace culture. *Journal of Pragmatics* 34, 1683–1710.

Hoover, David. L. 2001. Statistical stylistics and authorship attribution: an empirical investigation. *Literary and Linguistic Computing* 16(4), 421–444.

Hoover, David. L. 2002. Frequent word sequences and statistical stylistics. *Literary and Linguistic Computing* 17(2), 157–180.

Hoover, David. L. 2003. Multivariate analysis and the study of style variation. *Literary and Linguistic Computing* 18(4), 341–360

Hoover, David. L. 2004. Testing Burrows's Delta. *Literary and Linguistic Computing* (19)4, 453–475.

Hoover, David, L. 2009. Word frequency, statistical stylistics and authorship attribution. In Dawn Archer (ed.) *What's in a word list? Investigating Word Frequency and Keyword Extraction*. Farnham: Ashgate, 35–51.

Hopper, Paul, J. 1988. Emergent Grammar and the A Priori Grammar Postulate. In Deborah Tannen (ed.) *Linguistics in Context*: *Connecting Observation and Understanding*. Norwood, NJ: Ablex, 117–134.

Hopper, Paul, J. 1998. Emergent grammar. In Michael Tomasello (ed.) *The New Psychology of Language: Cognitive and Functional Approaches to Language Structure*. Mahwah, NJ: Erlbaum, 155–175.

Howald, Blake, Stephen. 2008. Authorship attribution under the rules of evidence: empirical approaches – a layperson's legal system. *The International Journal of Speech, Language and the Law* 15(2), 219–247.

Hoye, Leo. 2014. *Adverbs and Modality in English*. London: Routledge.

Hyde, Janet, Shibley. 2005. The gender similarities hypothesis. *American Psychologist* 60(6), 581–592.

Hymes, Dell. 1974. *Foundations in Sociolinguistics: An Ethnographic Approach*. Philadelphia: University of Pennsylvania Press.

IAFL. 2006. *The International Association of Forensic Linguists* website. [online]. Available from http://www.iafl.org/. [Accessed January 2010].

IAFL. 2013. *The International Association of Forensic Linguists* website – 'Constitution'. [online]. Available from: http://www.iafl.org/constitution.php. [Accessed August 2014].

IBM Corp. 2012. *IBM SPSS Statistics for Windows, Version 21.0*. Armonk, NY: IBM Corp.

Illocution Inc. 2013. *Enron Full Set*. [online]. Available from: http://www.illocutioninc.com/Corpora/. [Accessed July 2012].

Investopedia. 'bookout'. [online]. Available from: http://www.investopedia.com/terms/b/bookout.asp. [Accessed April 2014].

Iqbal, Farkhund. Rachid Hadjidj, Benjamin C.M. Fung, Mourad Debbabi. 2008. A novel approach of mining write-prints for authorship attribution in e-mail forensics. *Digital Investigation* 5(supplement), S42–S51.

Iqbal, Farkhund, Hamad Binsalleeh, Benjamin C.M. Fung, and Mourad Debbabi. 2010. Mining writeprints from anonymous e-mails for forensic investigation. *Digital Investigation* 7(1/2), 56–64.

Irvine, Judith. 2001.'Style' as distinctiveness: the culture and ideology of linguistic differentiation. In Penelope Eckert and John R. Rickford (eds.) *Style and Sociolinguistic Variation*. Cambridge: Cambridge University Press, 21–43.

Ishihara, Shunichi. 2014. A likelihood ratio-based evaluation of strength of authorship attribution evidence in SMS messages using N-grams. *The International Journal of Speech, Language and the Law* 21(1), 23–50.

Izsak, C. and Andrew R.G. Price. 2001 Measuring beta-diversity using a taxonomic similarity index, and its relation to spatial scale. *Marine Ecology Progress Series* 215, 69–77.

Jaccard, Paul. 1912. The distribution of the Wora in the alpine zone. *The New Phytologist* 11(2), 37–50.

Jay, Timothy. 1992. *Cursing in America: A Psycholinguistic Study of Dirty Language in the Courts, in the Movies, in the Schoolyards and on the Streets*. Amsterdam: John Benjamins.

Jeffries, Leslie and Brian Walker. 2012. Key words in the press: A critical corpus-driven analysis of ideology in the Blair years (1998- 2007). *English Text Construction* 5(2), 208–229.

Jockers, Matthew. L. and Daniela M. Witten. 2010. A comparative study of machine learning methods for authorship attribution. *Literary and Linguistic Computing* 25(2), 215–223.

Johnson, Alison and David Woolls. 2009. Who wrote this? The linguist as detective. In Susan Hunston and David Oakey (eds.) *Introducing Applied Linguistics: Concepts and Skills*. London: Routledge, 111–118.

Johnson, Alison and David Wright. 2014. Identifying idiolect in forensic authorship attribution: an n-gram textbite approach. *Language and Law (Linguagem e Direito)* 1(1), 37–69.

Johnson, Sally. 1997. Theorizing language and masculinity: a feminist perspective. In Sally Johnson and Ulrike Hanna Meinhof (eds.) *Language and Masculinity*. Oxford: Blackwell, 8–26.

Johnstone, Barbara. 1996. *The Linguistic Individual: Self Expression in Language and Linguistics*. Oxford: Oxford University Press.

Johnstone, Barbara. 2009 Stance, style, and the linguistic individual. In Alexander Jaffe (ed.) *Stance: Sociolinguistic Perspectives.* Oxford: Oxford University Press, 29–52.

Journal of Law and Policy. 2013. Volume 21, Issue 2. Special issue of proceedings from workshop on authorship attribution. (June 2013).

Juola, Patrick. 2008. *Authorship Attribution*. Delft: NOW Publishing.

Juola, Patrick. 2013. Stylometry and immigration: A case study. *Journal of Law and Policy* 21(2), 287–298.

Kaltenbock, Gunther. 2009. Initial *I think*: Main or comment clause? *Discourse and Interaction* 2(1), 49–70.

Kankaanranta, Anne. 2005. *''Hej Seppo, Could You Pls Comment on This!'' Internal Email Communication in Lingua Franca English in a Multinational Company*. PhD thesis. University of Jyväskylä. [online]. Available from: http://ebooks.jyu.fi/solki/9513923207.pdf. [Accessed January 2013].

Karkkainen, Elise. 2003. *Epistemic Stance in English Conversation. A Description of its Interactional Functions, with a Focus on 'I Think'*. Amsterdam: Benjamins.

Kecskés, Istvan. 2000. A cognitive-pragmatic approach to situation-bound utterances. *Journal of Pragmatics* 32(5), 605–625.

Kehoe, Andrew and Matt Gee. 2009. Weaving Web data into a diachronic corpus patchwork. In Antoinette Renouf and Andrew Kehoe (eds.) *Corpus Linguistics: Refinements and Reassessments.* Amsterdam: Rodopi, 225–279.

Kendall, Shari and Deborah Tannen. 1997. Gender and language in the workplace. In Ruth Wodak (ed.) *Gender and Discourse*. London: Sage, 81–105.

Kessler, Greg. 2010. Virtual business: An Enron email corpus study. *Journal of Pragmatics* 42, 262–270.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý and Vít Suchomel. 2014. The Sketch Engine: Ten years on. *Lexicography* 1(1), 1–30. [online]. Available from: http://www.sketchengine.co.uk.

Kjell, Bradley. 1994. Authorship determination using letter pair frequency features with neural network classifiers. *Literary and Linguistic Computing* 9(2), 119–124.

Kleifgen, Jo, Anne. 2001. Assembling talk: social alignments in the workplace. *Research on Language and Social Interaction* 34(3), 279–308.

Klimt, Bryan and Yiming Yang. 2004. Introducing the Enron Corpus. In *Proceedings of the First Conference on Email and Anti-Spam (CEAS).* [online]. Available from: http://ceas.cc/2004/168.pdf. [Accessed July 2012]

Koehler, Jonathon, J. 2013. Linguistic confusion in court: evidence from the forensic sciences. *Journal of Law and Policy* 21(2), 515–540.

Koester, Almut. 2004. Relational sequences in workplace genres. *Journal of Pragmatics* 36, 1405–1428.

Koester, Almut. 2006. *Investigating Workplace Discourse*. London: Routledge.

Koester, Almut. 2010a. Building small specialised corpora. In Anne O'Keefe and Michael McCarthy (eds.) *The Routledge Handbook of Corpus Linguistics.* London: Routledge, 66–79.

Koester, Almut. 2010b. *Workplace Discourse*. London: Continuum.

Koppel, Moshe, Argamon, Shlomo and Anat Rachel Shimoni. 2002. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing* 17(4), 401–412.

Koppel, Moshe and Jonathan Schler. 2003. Exploiting stylistic idiosyncrasies for authorship attribution. In *Proceedings of the 18th IJCAI Workshop on Computational Approaches to Style Analysis and Synthesis*. [online]. Available from: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.2.3019. [Accessed December 2010].

Koppel, Moshe, Jonathan Schler, Kfir Zigdon. 2005. Determining an author's native language by mining a text for errors. In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining (KDD '05)*, 624–628. [online]. Available from http://u.cs.biu.ac.il/~schlerj/schler_kdd05.pdf [Accessed January 2014].

Koppel, Moshe, Jonathan Schler and Shlomo Argamon. 2009. Computational methods in authorship attribution. *Journal of the American Society for Information Science And Technology* 60(1), 9–26.

Koppel, Moshe, Jonathan Schler, Shlomo Argamon. 2011. Authorship attribution in the wild. *Language Resources and Evaluation* 45(1), 83–94.

Koppel, Mosche, Jonathon Schler and Shlomo Argamon. 2013. Authorship attribution: What's easy and what's hard? *Journal of Law and Policy* 21(2), 317–332.

Kredens, Krzysztof. 2002. Towards a corpus-based methodology of forensic authorship attribution: a comparative study of two idiolects. In Barbara Lewandowska-Tomaszczyk (ed.) *PALC'01: Practical Applications in Language Corpora*. Peter Lang: Frankfurt am Mein, 405–437.

Kredens, Krzysztof. 2014. *Email to David Wright*. 10 July 2014.

Kredens, Krzysztof and Malcolm Coulthard. 2012. Corpus Linguistics in authorship identification. In Peter Tiersma and Lawrence M. Solan (eds.) *The Oxford Handbook of Language and Law*. Oxford: Oxford University Press, 504–516.

Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of HLT-NAACL 2003*, 252–259.

Kucukyilmaz, Tayfun, Barla Cambazogl, Cevdet Aykanat and Fazli Can. 2008. Chat mining: Predicting user and message attributes in computer-mediated communication. *Information Processing and Management* 44(4), 1448–1466.

Kuhl, Joseph, W. 2003. *The Idiolect, Chaos, and Language Custom Far From Equilibrium: Conversations in Morocco*. PhD Thesis, University of Georgia, Athens, Georgia. [online]. Available from: https://getd.libs.uga.edu/pdfs/kuhl_joe_w_200308_phd.pdf [Accessed March 2013].

Labbé, Cyril and Dominique Labbé. 2001. Inter-textual distance and authorship attribution Corneille and Moliére. *Journal of Quantitative Linguistics* 8(3), 213–231.

Labbé, Dominique. 2007. Experiments on authorship attribution by intertextual distance in English. *Journal of Quantitative Linguistics* 14(1), 33–80.

Labov, William. 1966. *The Social Stratification of English in New York City*. Washington, DC: Center for Applied Linguistics.

Labov, William. 1972. *Sociolinguistic Patterns*. Philadelphia : University of Pennsylvania.

Lampert, Andrew, Robert Dale and Cécile Paris. 2008. The nature of requests and commitments in email messages. In Mark Dredze, Vitor R. Carvalho and Tessa Lau (eds.) *Enhanced Messaging: Papers from the AAAI* Workshop. California: AAAI Press. [online]. Available from: http://clt.mq.edu.au/~rdale/teaching/itec810/2009H1/samples/AAAI2008-Lampert-RequestsAndCommitmentsInEmail.pdf [Accessed January 2013].

Larner, Samuel. 2014. A preliminary investigation into the use of fixed formulaic sequences as a marker of authorship. *The International Journal of Speech, Language and the Law* 21(1), 1–22.

Laver, John. 1981. Linguistic routines and politeness in greeting and parting. In Florian Coulmas (ed.) *Conversational Routine*. The Hague: Mouton, 289–304.

Leech, Geoffrey. 1992. Corpora and theories of linguistic performance. In Jan Svartvik (ed.) *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82*. Berlin: Mouton de Gruyter, 125–148.

Leech, Geoffrey, Roger Garside and Michael Bryant. 1994. CLAWS4: the tagging of the British National Corpus. In *Proceedings of the 15th International Conference on Computational Linguistics*, 622–628.

Leech, Geoffrey, Nicholas Smith and Paul Rayson. 2012. English style on the move: Variation and change in stylistic norms in the twentieth century. In Merja Kytö *English Corpus Linguistics: Crossing Paths*. Amsterdam: Rodopi, 69–98.

Li, Min, Youngja Park, Rui Ma, He, Y. Huang. 2012. Business email classification using incremental subspace learning. Paper presented at *21st International Conference on Pattern Recognition*. Tsukuba, Japan, November 2012.

Louw, Bill. 1993. Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In Mona Baker, Gill Francis and Elena Tognini-Bonelli (eds.) *Text and Technology: In Honour of John Sinclair*. Amsterdam/Philadelphia: John Benjamins, 157–176.

Love, Harold. 2002. *Attributing Authorship: An Introduction*. Cambridge: Cambridge University Press.

Lubarski, Pawel and Mikołaj Morzy. 2012. Measuring the importance of users in a social network based on email communication patterns. In *Proceedings of the 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining,* (*ASONAM*). [online]. Available from: http://dataminingalapolonaise.files.wordpress.com/2012/08/asonam2012.pdf. [Accessed March 2013].

Lucy, David. 2005. *Introduction to Statistics for Forensic Scientists*. London: Wiley.

Luyckx, Kim and Walter Daelemans. 2008. Using syntactic features to predict author personality from text. In *Proceedings of Digital Humanities 2008,* 146–149. [online]. Available from: http://www.cnts.ua.ac.be/papers/2008/LD08dh.pdf. [Accessed January 2013].

Luyckx, Kim and Daelemans, Walter. 2011. The effect of author set size and data size in authorship attribution. *Literary and Linguistic Computing* 26(1), 35–55.

Mackey, Alison and Susan M. Gass. 2005. *Second Language Research: Methodology and Design*. London: Routledge.

MacLeod, Nicci and Tim Grant. 2012. Whose Tweet? Authorship analysis of micro-blogs and other short form messages. In Sam Tomblin, Nicci MacLeod, Rui Sousa-Silva and Malcolm Coulthard (eds.) *Proceedings of the Tenth International Association of Forensic Linguists' Biennial Conference,* Aston University, Birmingham, 210–224. [online]. Available from: www.forensiclinguistics.net. [Accessed December 2012].

Maes, Freek and Johannes C. Scholtes. 2012. Authorship disambiguation and alias resolution in email data. Paper presented at *24th Benelux Conference on Artificial Intelligence*. Maastricht, Netherlands, October 2012.

Mahlberg, Michaela and Dan McIntyre. 2011. A case for corpus stylistics: Ian Fleming's Casino Royale. *English Text Construction* 4(2), 204–227.

Maley, Yon and Rhondda Fahey. 1991. Presenting the evidence**:** Constructions of reality in court. *International Journal for the Semiotics of Law* 4(1), 3–17.

Manasse, Mark, S. 2012. *On the Efficient Determination of Most Near Neighbors: Horseshoes, Hand Grenades, Web Search, and other Situations When Close is Close Enough.* San Rafael: Morgan and Claypool.

Manca, Elena. 2010. From phraseology to culture: Qualifying adjectives in the language of tourism. In Ute Römer and Rainer Schulze (eds.) *Patterns, Meaningful Units and Specialized Discourses*. Amsterdam: John Benjamins, 105–122.

Markson, Lucy, Jessica Woodhams and John W. Bond. 2010. Linking serial residential burglary: comparing the utility of *modus operandi* behaviours, geographical proximity, and temporal proximity. *Journal of Investigative Psychology and Offender Profiling* 7(2), 91–107.

McCallum, Andrew, Xuerui Wang and Andrés Corrada-Emmanuel. 2007. Topic and role discovery in social networks with experiments on Enron and academic Email. *Journal of Artificial Intelligence Research* 30(3), 249–272.

McCarthy, Michael. 2000. Mutually captive audiences: small talk and the genre of close-contact service encounters. In Justine Coupland (ed.) *Small Talk.* London: Routledge, 84–109.

McColly, William, B. and Dennis Weier. 1983. Literary attribution and likelihood-ratio tests: the Case of the Middle English Pearl-poems. *Computers and the Humanities* 17(2), 65 –75.

McCrae, Robert, R. and Paul T. Costa Jr. 1996. Toward a new generation of personality theories: Theoretical contexts for the five-factor model. In Jerry S.Wiggins (ed.) *The Five-Factor Model of Personality: Theoretical Perspectives*. Guilford: New York, 51–87.

McEnery, Tony and Andrew Wilson. 2001. *Corpus Linguistics: An Introduction* (2nd edn.) Edinburgh: Edinburgh University Press.

McEnery, Tony, Richard Xiao and Yukio Tono. 2006. *Corpus-based language studies : an advanced resource book.* London: Routledge.

McEnery, Tony and Andrew Hardie. 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

McMenamin, Gerald .R. 2002 *Forensic Linguistics: Advances in Forensic Stylistics*. Boca Raton, Florida: CRC Press.

McMenamin, Gerald. R. 2004. Disputed authorship in US law. *International Journal of Speech, Language and the Law* 11(1), 74–82.

McMenamin, Gerald R. 2010. Forensic stylistics. Theory and practice of forensic stylistics. In Malcolm Coulthard and Alison Johnson (eds.) *The Routledge Handbook of Forensic Linguistics*. London: Routledge, 487–507

Michelbacher, Lukas, Stefan Evert, and Hinrich Schutze. 2011. Asymmetry in corpus-derived and human word associations. *Corpus Linguistics and Linguistic Theory* 7(2), 245–276.

Milroy, James and Lesley Milroy. 1978. Belfast: change and variation in an urban vernacular. In Peter Trudgill (ed.) *Sociolinguistic Patterns in British English*. London: Arnold, 19–36.

Mitra, Tanushree and Eric Gilbert. 2012. Have you heard?: How gossip flows through workplace email. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media.* [online]. Available from: http://comp.social.gatech.edu/papers/icwsm12.gossip.mitra.pdf . [Accessed September 2013].

Mollin, Sandra. 2009a. Combining corpus linguistic and psychological data on word co-occurrences: corpus collocates versus word associations. *Corpus Linguistics and Linguistic Theory* 5(2), 175 –200.

Mollin, Sandra. 2009b. 'I entirely understand' is a Blairism: The methodology of identifying idiolectal collocations. *International Journal of Corpus Linguistics* 14(3), 367–392.

Moore, Roger, K. 2001. There's no data like more data (but when will enough be enough?). In *Proceedings of the Institute of Acoustics Workshop on Innovation in Speech Processing* 23(3), 19–26.

Mosteller, Frederick. and Wallace, David. L. 1964. *Inference and Disputed Authorship: The Federalist*. Reading, MA: Addison-Wesley Publishing Company Inc.

Murphy, Brona. 2010. *Corpus and Sociolinguistics: Investigating Age and Gender in Female Talk.* Amsterdam: John Benjamins.

Murphy, Brona and Dawn Knight. 2014. Exploring the meta in 'meta-data': corpus investigations in sociolinguistic contexts. Paper presented at *The Seventh Inter-Varietal Applied Corpus Studies (IVACS) International Conference*, University of Newcastle, 19–21 June, 2014.

Narayanan, Arvind, Hristo Paskov, Neil Zhenqiang Gong, John Bethencourt, Emil Stefanov, Eui Chul Richard Shin and Dawn Song. 2012. On the Feasibility of Internet-Scale Author Identification. Paper presented at *IEEE Security and Privacy symposium 2012*. San Francisco, USA, May 2012.

Nattinger, James R. and Jeanette DeCarrico. 1992. *Lexical Phrases and Language Teaching*. Oxford: Oxford University Press.

Naumann, Felix Melanie Herschel. 2010. An introduction to duplicate detection. *Synthesis Lectures on Data Management* 2(1), 1–87.

Nerlich, Brigitte, Richard Forsyth, and David Clarke. 2012. Climate in the news: How differences in media discourse between the US and UK reflect national priorities. *Environmental Communication: A Journal of Nature and Culture* 6(1), 44–63.

Neumann, Hendrik and Martin Schnurrenberger. 2009. E-Mail authorship attribution applied to the Extended Enron Authorship Corpus (XEAC). [online]. Available from: http://code.google.com/p/eyebachelor/downloads/list. [Accessed December 2010].

New York Times, The. 2006. 10 Enron players: Where they landed after the fall. 29 January 2009. [online]. Available from: http://www.nytimes.com/2006/01/29/business/businessspecial3/29profiles.html?adxnnl=1&pagewanted=all&adxnnlx=1410973975-3O0EJUlWFjMiZ2nS/UXDRg. [Accessed July 2014].

Nini, Andrea and Tim Grant. 2013. Bridging the gap between stylistic and cognitive approaches to authorship analysis using Systemic Functional Linguistics and multidimensional analysis. *The International Journal of Speech, Language and the Law* 20(2), 173–202.

Noecker, John, Michael Ryan and Patrick Juola. 2013. Psychological profiling through textual analysis. *Literary and Linguistic Computing* 28(3), 382–387.

Oakes, Michael. 1998. *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.

O'Halloran, Kieran, A. 2009. Inferencing and cultural reproduction: a corpus-based critical discourse analysis. *Text and Talk* 29(1), 21–52.

O'Keeffe, Anne, Michael McCarthy and Ronald Carter. 2007. *From Corpus to Classroom: Language Use and Language Teaching*. Cambridge: Cambridge University Press.

Omerod, David. 1999. Criminal profiling: Trial by judge and jury, not criminal psychologist. In David Canter and Laurence Alison (eds.) *Profiling in Policy and Practice*. Aldershot: Ashgate.

Omerod, David and Jim Sturman. 2005. Working with the courts: Advice for expert witnesses. In Laurence Alison (ed.) *The Forensic Psychologists' Casebook: Psychological Profiling and Criminal Investigation*. London: Routledge, 170–193.

Oxford English Dictionary, The. 'deal, v.' and 'deal, n.2'. Oxford University Press. [online]. Available from: www.oed.com. [Accessed July 2014].

Pace-Sigge, Michael. 2013. *Lexical Priming in Spoken English Usage*. London: Palgrave.

Palus, Sebastian, Piotr Bródka and Przemysław Kazienko. 2010. How to analyze company using social network? In Miltiadis D. Lytras, Patricia Ordonez De Pablos, Adrian Ziderman, Alan Roulstone, Hermann Maurer and Jonathan B. Imber (eds.) *Knowledge Management, Information Systems, E-Learning, and Sustainability Research*. Berlin: Springer, 159–164.

Paul, Hermann. 1888. *Principles of the History of Language*. (Translated by H.A Strong). London: Swan, Sonnenschien, Lowrey and Co.

Paulsen, Derek. 2006. Human versus machine: A comparison of the accuracy of geographic profiling methods. *Journal of Investigative Psychology and Offender Profiling* 3(2), 77–89.

Pennebaker, James W., Matthias R. Mehl, and Kate G. Niederhoffer. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology* 54, 547–577.

Phakiti, Aek. 2010. Analysing Quantitative Data. In Brian Paltridge and Aek Phakiti (eds.) *Continuum Companion to Research Methods in Applied Linguistics.* London: Continuum, 39–49.

Pham, Dang Duc Giang Binh Tran, Son Bao Pham. 2009. Author Profiling for Vietnamese Blogs. In *Proceedings of the 2009 International Conference on Asian Language Processing*, 190–194.

Pierrehumbert, Janet, Breckenridge. 2001. Exemplar dynamics: Word frequency, lenition and contrast. In Joan, L. Bybee and Paul J. Hopper (eds.) *Frequency and the Emergence of Linguistic Structure*. Amsterdam: John Benjamins, 137–157.

Planken, Brigitte. 2005. Managing rapport in lingua franca sales negotiations: A comparison of professional and aspiring negotiators. *English for Specific Purposes* 24(4), 381 400.

Poncini, Gina. 2004. *Discursive Strategies in Multicultural Business Meetings*. Bern: Peter Lang.

Pottier, Julien, Anne Dubuis, Loïc Pellissier, Luigi Maiorano, Leila Rossier, Christophe F. Randin, Pascal Vittoz, and Antoine Guisan. 2013. The accuracy of plant assemblage prediction from species distribution models varies along environmental gradients. *Global Ecology and Biogeography* 22(1), 52–63.

Potts, Amanda and Paul Baker. 2012. Does semantic tagging identify cultural change in British and American English? *International Journal of Corpus Linguistics* 17(3), 295–324.

Powers, William C., Raymond S. Troubh and Herbert S. Winokur. 2002. Report of Investigation by the Special Investigative Committee of the Board of Directors of Enron Corp. Diane Publishing Company.

Priebe, Carey, E., John M. Conroy, David J. Marchette, Youngster Park. 2005. *Scan Statistics on Enron Graphs. Computational and Mathematical Organization Theory* 11(3), 229–247.

Queralt, Sheila and M. Teresa Turell. 2012. Testing the discriminatory potential of sequences of linguistic categories (n-grams) in Spanish, Catalan and English corpora. Paper presented at *the Regional Conference of the International Association of Forensic Linguists*. University of Malaysia, Kuala Lumpur, 4–7 July 2012.

Radicati Group Inc. 2013. *Email Statistics Report 2013 –2017.* [online]. Available from: http://www.radicati.com/wp/wp-content/uploads/2013/04/Email-Statistics-Report-2013-2017-Executive-Summary.pdf. [Accessed November 2013].

Rajaraman, Anand and Jeffrey David Ullman. 2011. *Mining of Massive Datasets*. Cambridge: Cambridge University Press.

Rangel, Francisco and Paolo Rosso. 2013. Use of language and author profiling: Identification of gender and age. In *Proceedings of the 10th International Workshop on Natural Language Processing and Cognitive Sciences*. [online]. Available from http://users.dsic.upv.es/~prosso/resources/RangelRosso_NLPCS13.pdf. [Accessed February 2014].

Rangel, Francisco, and Paolo Rosso, Moshe Koppel, Efstathios Stamatatos and Giacomo Inches. 2013. Overview of the author profiling task at PAN 2013. *Working Notes Papers of the CLEF 2013 Evaluation Labs* [online]. Available from: http://www.uni-weimar.de/medien/webis/research/events/pan-13/pan13-web/about.html. [Accessed December 2013].

Rashid, Awais Alistair Baron, Paul Rayson, and Corinne May-Chahal. 2013. Who am I? Analyzing digital personas in cybercrime investigations. *Computer* 46(4), 54–61.

Rayson, Paul. 2003. *Matrix: A Statistical Method and Software Tool for Linguistic Analysis through Corpus Comparison*. PhD Thesis, Lancaster University.

Rayson, Paul and Roger Garside. 1998. The CLAWS Web Tagger. *ICAME Journal* 22, 121–123.

Rayson, Paul and Roger Garside. 2000. Comparing corpora using frequency profiling. In *Proceedings of the workshop on Comparing Corpora* (held in conjunction with the 38th annual meeting of the Association for Computational Linguistics), 1–6.

Rayson, Paul, Damon Berridge and Brian Francis. 2004. Extending the Cochran rule for the comparison of word frequencies between corpora. In *Proceedings of the 7th International Conference on Statistical Analysis of Textual Data (JADT 2004)*, 926–936. [online]. Available from: http://www.comp.lancs.ac.uk/~paul/publications/rbf04_jadt.pdf [Accessed May 2014] .

Renouf, Antoinette and John, M. Sinclair. 1991. Collocational frameworks in English. In Karin Aijmer and Bengt Altenberg (eds.) *English corpus linguistics*. New York: Longman, 128–143.

Rico-Sulayes, Antonio. 2011 Statistical authorship attribution of Mexican drug trafficking online forum posts. *The International Journal of Speech, Language and the Law* 18(1), 53–74.

Rose, Phil. 2002. *Forensic speaker identification*. London: Taylor and Francis.

Rose, Phil. 2006. Technical forensic speaker recognition: Evaluation, types and testing of evidence. *Computer Speech and Language* 20(2/3), 159–191.

Rose, Phil. 2013. More is better: Likelihood ratio-based forensic voice comparison with vocalic segmental cepstra frontends. *The International Journal of Speech, Language and the Law* 20(1), 77–116.

Rungruangthum, Montarat, Richard Watson Todd and Wirote Aroonmanakun. 2011. Re-conceptualization of piloting research in applied linguistics. In *Proceedings of the International Conference: Doing Research in Applied Linguistics***,** 27–33. [online]. Available from: http://arts.kmutt.ac.th/dral/index.php?q=content/proceedings-international-conference [Accessed May 2012].

Ryan Rowe, Germán Creamer, Shlomo Hershkop and Salvatore, J, Stolfo. 2007. Automated social hierarchy detection through email network analysis. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*. New York: ACM, 109–117. [online]. Available from: academiccommons.columbia.edu/download/fedora.../cucs-040-07.pdf. [Accessed April 2012].

Sairio, Anni. 2006. Progressives in the letters of Mrs. Elizabeth Montagu and her circle in 1738–1778. In Christiane Dalton-Puffer, Nikolaus Ritt, Herbert Schendl, and Dieter Kastovsky (eds.) *Syntax, Style and Grammatical Norms: English from 1500-2000*. *Linguistic Insights*. Frankfurt: Peter Lang, 167–190.

Sairio, Anni. 2014. The linguistic biography of Elizabeth Montagu (1718–1800): Aims, methods, and approaches. Paper presented at *Sociolinguistics Symposium 20*, University of Jyväskylä, Finland, 15–18 June 2014.

Sanderson, Conrad and Simon Guenter. 2006. Short text authorship attribution via sequence kernels, Markov chains and author unmasking: An investigation. In *Proceedings of the International Conference on Empirical Methods in Natural Language Engineering*. Morristown, NJ: Association for Computational Linguistics, 482–491.

Sapir, Edward. 1927. Speech as a personality trait. *American Journal of Sociology* 32(6), 892–905.

Saussure, Ferdinand de. 1966 [1916]. *Course in general linguistics*. Edited by Charles Bally, and Albert Sechehaye. Translated by Wade Baskin. New York: McGraw-Hill. (Original work published in 1916).

Savoy, Jacques. 2012. Authorship attribution: A comparative study of three text corpora and three languages. *Journal of Quantitative Linguistics* 19(2), 132–161.

Savoy, Jacques. 2013. Authorship attribution based on a probabilistic topic model. *Information and Processing Management* 49(1), 341–354.

Schler, Jonathan, Mosche Koppel, Shlomo Argamon and James W. Pennebaker. 2006. Effects of age and gender on Blogging. In *Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*. [online]. Available from: http://u.cs.biu.ac.il/~schlerj/schler_springsymp06.pdf [Accessed February 2013].

Schmitt, Norbert and Ronald Carter. 2004. Formulaic sequences in action: An introduction. In Norbert Schmitt (ed.) *Formulaic Sequences: Acquisition, Processing and Use.* Amsterdam/Philadelphia : John Benjamins, 1–22.

Schmitt, Norbert, Sarah Grandage, and Svenja Adolphs. 2004. Are corpus-derived recurrent clusters psycholinguistically valid? In Norbert Schmitt (ed.) *Formulaic Sequences: Acquisition, Processing and Use.* Amsterdam: John Benjamins, 127–151.

Schneider, Klaus P. 2008. Small talk in England, Ireland, and the USA. In Klaus P. Schneider and Anne Barron (eds.) *Variational Pragmatics: A Focus on Regional Varieties in Pluricentric Languages*, 99–139.

Schneidman, Edwin, S. and Norman L. Farberow (eds.) 1957. *Clues to Suicide*. New York: McGraw Hill.

SciVerse Scopus. 2013. Elsevier. [online]. Available from: http://www.scopus.com/home.url [Accessed August 2014].

Scott, Mike and Tim Johns. 1993. *MicroConcord*. Oxford: Oxford University Press.

Scott, Mike and Christopher Tribble. 2006. *Textual Patterns: Keyword and Corpus Analysis in Language Education*. Amsterdam: John Benjamins.

Scott, Mike. 2008a. *WordSmith Tools version 5*. Liverpool: Lexical Analysis Software.

Scott, Mike.2008b. *WordSmith Tools Help*. Liverpool: Lexical Analysis Software.

Searle, John, R. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge: Cambridge University Press.

Shapero, Jess. 2011. *The Language of Suicide Notes*. PhD thesis, University of Birmingham. [online]. Available from: http://etheses.bham.ac.uk/1525/. [Accessed May 2013].

Sinclair, John. M. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, John, M. 1996. The search for units of meaning. *Textus* 9(1), 71–106.

Sinclair, John, M. 2004. *How to Use Corpora in Language Teaching*. Amsterdam: John Benjamins.

Snook, Brent, Richard M. Cullen, Craig Bennell, Paul J. Taylor and Paul Gendreau. 2008. The criminal profiling illusion: What's behind the smoke and mirrors? *Criminal Justice and Behavior* 35(10), 1257–1276.

Solan, Lawrence, M. 2013. Intuition versus algorithm: The case for forensic authorship attribution. *Journal of Law and Policy* 21(2), 551–576.

Solan, Lawrence, M. and Peter M. Tiersma. 2004. Author Identification in American Courts. *Applied Linguistics* 25(4), 448–465.

Stamatatos, Efstathios. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology* 60(3), 538–556.

Stamatatos, Efstathios. 2013. On the robustness of authorship attribution based on character n-gram features. *Journal of Law and Policy* 21(2), 421–440.

Stamatatos, Efstathios, Nikos Fakotakis, Georgios Kokkinakis. 2001. Computer-based authorship attribution without lexical measures. *Computers and the Humanities 35*(2), 193–214.

Stein, Mark and Jonathan Pinto. 2011. The dark side of groups: A 'gang at work' in Enron. *Group and Organization Management* 36(6), 692–721.

Stubbs, Michael. 2001. *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford: Blackwell.

Styler, Will. 2011. *The EnronSent Corpus*. Technical Report 01-2011, University of Colorado at Boulder Institute of Cognitive Science, Boulder, CO. [online]. Available from: http://verbs.colorado.edu/enronsent/ [Accessed April 2013].

Svartvik, Jan. 1968. *The Evans Statements: A case for Forensic Linguistics*. Gotëborg: University of Gothenburg Press.

Tagg, Caroline. 2009. *A Corpus Linguistics Study of SMS Text Messaging*. PhD thesis, University of Birmingham. [online]. Available from: http://etheses.bham.ac.uk/253/1/Tagg09PhD.pdf [Accessed January 2013].

Tagliamonte, Sali A., Mercedes Durham and Jennifer Smith. 2014. Grammaticalization at an early stage: future *be going to* in conservative British dialects. *English Language and Linguistics* 18(1), 75–108.

Talbot, Mary. 2010. *Language and Gender* (2nd edn.). Cambridge: Polity.

Tallentire, David, R. 1973. Towards an archive of lexical norms – a proposal. In Adam, J. Aitken, Richard, W. Bailey and Neil Hamilton-Smith (eds.) *The Computer and Literary Studies*. Edinburgh: Edinburgh University Press, 39–60.

Tallentire, David, R. 1976. Confirming intuitions about style using concordances. In Alan Jones and Robert, F. Churchouse (eds.) *The Computer in Literary and Linguistic Studies.* Cardiff: University of Wales Press, 309–328.

Tam, Tony, Artur J. Ferreira, André Lourenço. 2012. Automatic foldering of email messages: a combination approach. Paper presented at *34th European Conference on Information Retrieval*. Barcelona, Spain, April 2012.

Tang, Zhiyao, Jingyun Fang, Xiulian Chi, Jianmeng Feng, Yining Liu, Zehao Shen, Xiangping Wang, Zhihen Wang, Xiaopu Wu, Chengyang Zheng, Kevin J. Gaston. 2013. Patterns of plant beta-diversity along elevational and latitudinal gradients in mountain forests of China. *Ecography* 35(12), 1083–1091.

Tannen, Deborah. 1990. Gender differences in topical coherence: creating involvement in best friends' talk. *Discourse Processes* 13(1), 73–90.

Taylor, Charlotte. 2013. Searching for similarity using corpus-assisted discourse studies. *Corpora* 8(1), 81–113.

Telegraph, The. 2013a. Husband killed and burnt wife, then took world trip on her cash. 11 May 2013. [online]. Available from: http://www.telegraph.co.uk/news/uknews/crime/10050579/Husband-killed-and-burnt-wife-then-took-world-trip-on-her-cash.html [Accessed August 2014].

Telegraph, The. 2013b. JK Rowling unmasked: the lawyer, the wife, her tweet - and a furious author. 21 July 2013. [online]. Available from: http://www.telegraph.co.uk/culture/books/10192275/JK-Rowling-unmasked-the-lawyer-the-wife-her-tweet-and-a-furious-author.html [Accessed August 2014].

Telegraph, The. 2014. Linguistic researchers begin hunt for the next 'selfie'. 3 March 2014. [online]. Available from: http://www.telegraph.co.uk/technology/twitter/10672643/Linguistic-researchers-begin-hunt-for-the-next-selfie.html [Accessed August 2014].

Thomson, Rob and Tamar Murachver. 2001. Predicting gender from electronic discourse. *British Journal of Social Psychology* 40(2), 193–208.

Tiersma, Peter and Lawrence M. Solan. 2002. The linguist on the witness stand: forensic linguistics in American courts. *Language* 78(2), 221–239.

Titak, Ashley and Audrey Roberson. 2013. Dimensions of web registers: an exploratory multi-dimensional comparison. *Corpora* 8(2), 235–260.

Tognini-Bonelli, Elena. 2001. *Corpus Linguistics at Work*. Amsterdam: John Benjamins.

Tomblin, Samuel. 2013. *To Cut a Long Story Short: An Analysis of Formulaic Sequences in Short Written Narratives and their Potential as Markers of Authorship*. PhD Thesis, Aston University, Birmingham. [online] Available from: http://eprints.aston.ac.uk/19268/ [Accessed December 2013].

Tonkin, Matthew, Tim Grant and John W. Bond. 2008. To link or not to link: A test of the case linkage principles using serial car theft data. *Journal of Investigative Psychology and Offender Profiling* 5(1/2), 59–77.

Toutanova, Kristina, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *Proceedings of HLT-NAACL 2003*, 252-259.

Tracy, Karen and Jessica Robles. 2009. Questions, questioning, and institutional practices: An introduction. *Discourse Studies* 11(2), 131–152.

Trudgill, Peter. 1972. Sex, covert prestige, and linguistic change in the urban British English of Norwich. *Language in Society* 1(2), 179–196.

Turell, M. Teresa. 2010. The use of textual, grammatical and sociolinguistic evidence in forensic text comparison. *The International Journal of Speech, Language and the Law* 17(2), 211–250.

Turell, M. Teresa. and Núria Gavaldà. 2013. Towards an index of idiolectal similitude (or distance) in forensic authorship analysis. *Journal of Law and Policy* 21(2), 495–514.

Turvey, Brent. 2012. *Criminal Profiling: An Introduction to Behavioral Evidence Analysis* (4th edn). Oxford: Elsevier.

Tweedie, Fiona. J., S. Singh, and David, I. Holmes. 1996. Neural network applications in stylometry: The Federalist Papers. *Computers and the Humanities* 30(1), 1–10.

van Den Eynden, Nadine. 2012. Politeness and gender in Belgian organisational emails. In Paul Gillaerts, Elizabeth de Groot, Sylvain Dieltjens, Priscilla Heynderickx and Geert Jacobs (eds.) *Researching Discourse in Business Genres: Cases and corpora.* Bern: Peter Lang, 33–52.

van Halteren, Hans Harald Baayen, Fiona Tweedie, Marco Haverkort and Anneke Neijt. 2005. New machine learning methods demonstrate the existence of a human Stylome. *Journal of Quantitative Linguistics* 12(1), 65–77.

van Halteren, Hans. 2008. Source language markers in Europarl translations. In *Proceedings of COLING2008, 22nd International Conference on Computational Linguistics*, 937– 944. [online]. Available from: http://www.aclweb.org/anthology/C08-1118 [Accessed January 2014].

Vine, Bernadette. 2004. *Getting Things Done at Work: The Discourse of Power in Workplace Interaction*. Amsterdam: John Benjamins.

Wachi, Taeko, Kazumi Watanabe, Kaeko Yokota, Mamoru Suzuki, Maki Hoshino, Atsushi Sato And Goro Fujita. 2007. Offender and crime characteristics of female serial arsonists in Japan. *Journal of Investigative Psychology and Offender Profiling* 4(1), 29–52.

Waldvogel, Joan. 2007 Greetings and closings in workplace email. *Journal of Computer-Mediated Communication* 12(2), 456–477.

Wales, Katie. 2001. *A Dictionary of Stylistics* (2nd edn.). London: Longman.

Walsh, Michael, Bernd Möbius, Travis Wade, Hinrich Schütze. 2010. Multilevel Exemplar Theory. *Cognitive Science* 34(4), 537–582.

Wang, Jinjun. 2006. Questions and the exercise of power. *Discourse and Society* 17(4), 529–548.

Wang, John. 2009. *EnronData.org: The Enron Data Reconstruction Project.* [online]. Available from: www.enrondata.org. [Accessed April 2013].

Wang, Man, Yifan He, and Minghu Jiang. 2010. Text categorization of Enron email corpus based on information bottleneck and maximal entropy. Paper presented at *IEEE 10th International Conference on Signal Processing*. Beijing, China, October 2010.

Weatherall, Ann. 2002. *Gender, Language and Discourse*. London: Routledge.

Webopedia. 'redlining'. [online]. Available from: http://www.webopedia.com/TERM/R/redlining.html. [Accessed July 2014].

Williams, Christopher. 2013. Changes in the verb phrase in legislative language in English. In Bas Aarts, Joanne Close, Geoffrey Leech and Sean. A. Wallis (eds.) *The Verb Phrase in English: Investigating Recent Language Change with Corpora.* Cambridge: Cambridge University Press, 353–371.

Williams, Graham. 2010. 'I haue trobled wth a tedious discours': Sincerity, sarcasm and seriousness in the letters of Maria Thynne, c. 1601-1610'. *Journal of Historical Pragmatics* 11(2), 169–193.

Williams, Graham. 2014. Written like a 'gwd' scotswoman: Margaret Tudor and English at the turn of the sixteenth century. Paper presented at *Sociolinguistics Symposium 20*, University of Jyväskylä, Finland, 15–18 June 2014.

Winter, Eugene. 1996. The statistics of analysing very short texts in a criminal context. In Hannes Kniffka (ed.) *Recent Developments in Forensic Linguistics.* New York: Peter Lang, 141–179.

Wong, Sze-Meng Jojo and Mark Dras. 2011. Exploiting parse structures for native language identification. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 1600–1610. [online]. Available from: http://www.aclweb.org/anthology/D11-1148 [Accessed January 2014].

Woodhams, Jessica and Kirsty Toye. 2007. An empirical test of the assumptions of case linkage and offender profiling with serial commercial robberies. *Psychology, Public Policy, and Law* 13(1), 59–85.

Woodhams, Jessica, Tim Grant and Andrew R.G. Price. 2007 From marine ecology to crime analysis: improving the detection of serial sexual offences using a taxonomic similarity measure. *The Journal of Investigative Psychology and Offender Profiling* 4(1), 17–27.

Woodhams, Jessica, Clive Hollin and Ray Bull. 2008. Incorporating context in linking crimes: an exploratory study of situational similarity and if-then contingencies. *Journal of Investigative Psychology and Offender Profiling* 5(1), 1–23.

Woods, Nicola. 1988. Talking shop: sex and status as determinants of floor apportionment in a work setting. In Jennifer Coates and Deborah Cameron (eds.) *Women in their Speech Communities*. London: Longman, 141–157.

Woolls, David. 2013. *CFL Jaccard n-gram Lexical Evaluator* (*Jangle*) version 2. CFL Software Limited.

Wray, Alison. 2000. Formulaic sequences in second language teaching: Principle and Practice. *Applied Linguistics* 21(4), 463–489.

Wray, Alison. 2002. *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.

Wray, Alison. 2008. *Formulaic Language: Pushing the Boundaries*. Oxford: Oxford University Press.

Wright, David. 2012. Existing and innovative techniques in authorship analysis: Evaluating and experimenting with computational approaches to 'big data' in the Enron Email Corpus. Paper presented at *The 3rd European Conference of the International Association of Forensic Linguists*, University of Porto, Portugal, 15–18 October 2012.

Wright, David. 2013. Stylistic variation within genre conventions in the Enron email corpus: Developing a text-sensitive methodology for authorship research. *International Journal of Speech, Language and the Law*, 20(1), 45–75.

Youngs, Donna. 2004. Personality correlates of offence style. *Journal of Investigative Psychology and Offender Profiling* 1(2), 99–119.

YouTube. 2012. *Vincent Kaminski*. Video posted by 'FEEMchannel' (Fondazione Eni Enrico Mattei). [online]. Available from: http://www.youtube.com/watch?v=byU9NPRqyLM. [Accessed July 2014].

Zaitsu, Wataru. 2010. Bomb threats and offender characteristics in Japan. *Journal of Investigative Psychology and Offender Profiling* 7(1), 75–89.

Zhang, Grace. 2014. The elasticity of *I think*: Stretching its pragmatic functions. *Intercultural Pragmatics* 11(2), 225–257.

Zheng, Rong Yi Qin, Zan Huang, and Hsinchun Chen. 2003. Authorship analysis in cybercrime investigation. In *Proceedings of the 1st NSF/NIJ Conference on Intelligence and Security Informatics*, 59–73.

Zheng, Rong, Jlexum Li, Hslnchun Chen and Zan Huang. 2006. A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American Society for Information Science and Technology* 57(3), 378–393.

Appendix 1

Technical Note: Description of CFL extraction routines for the CFL Enron Sent email database.

The above dataset used in the thesis had been largely created from the Carnegie-Mellon University (CMU) release of the Enron email database prior to the commencement of the thesis. Some amendments were made to the dataset during the thesis, and these are detailed at the end of this document. The central purpose of creating this subset was for the exploration of authorship attribution, so the only the folders named either **sent**, **sent_items** or **_sent_items**, which depended on either user preference or the original email product, were used. Where all these three potential types are in use, the folder with the most entries has been selected in all cases, to exclude duplicates.

This is an example of how each message in the CMU is identified: **kean-s/sent_items/122**. This has a unique email **custodian name** , then the sub folder for each custodian and finally the **sequential number** of the extracted message. This sequential number is not the actual message reference; it reflects the order in which the extraction program used by CMU processed the contents of each email folder, so each folder contents runs from 1 to the number of emails in that folder. The extracted set contains a large number of sub-folders, including Inbox and any named topic folders created by the custodian.

Message-ID:<16170877.1075858881106.JavaMail.evans@thyme>¶
Date:·Wed,·10·Oct·2001·11:49:31·-0700·(PDT)¶
From:·j..kean@enron.com¶
To:·jean.ryall@enron.com¶
Subject:·FW:·Wallis·Jefferson¶
Cc:·sherri.sera@enron.com¶
Mime-Version:·1.0¶
Content-Type:·text/plain;·charset=us-ascii¶
Content-Transfer-Encoding:·7bit¶
Bcc:·sherri.sera@enron.com¶
X-From:·Kean,·Steven·J.·</O=ENRON/OU=NA/CN=RECIPIENTS/CN=SKEAN>¶
X-To:·Ryall,·Jean·</O=ENRON/OU=NA/CN=RECIPIENTS/CN=Jryall>¶
X-cc:·Sera,·Sherri·</O=ENRON/OU=NA/CN=RECIPIENTS/CN=Notesaddr/cn=37d55f10-3d55d1ab-86256ada-71156a>¶
X-bcc:·¶
X-Folder:·\SKEAN·(Non-Privileged)\Kean,·Steven·J.\Sent·Items¶
X-Origin:·Kean-S¶
X-FileName:·SKEAN·(Non-Privileged).pst¶
¶
What·do·you·know·about·this·guy?¶
-----Original·Message-----¶
From:·Sherri·Sera·[mailto:sherri_sera@hotmail.com]¶
Sent:·Wednesday,·October·10,·2001·1:47·PM¶
To:·Kean,·Steven·J.¶
Subject:·Wallis·Jefferson¶
¶
Hi,·Steve.·I·hope·all·is·well.¶
¶
Jeff·wanted·me·to·ask·you·about·Wallis·Jefferson.··Apparently,·he's·running·¶
for·Texas·Supreme·Court·and·he·wanted·to·know·if·this·guy·"lives·the·¶
values,"·so·to·speak.··He·hasn't·called·him·yet,·but·just·wanted·your·¶
perspective·on·him·when·you·have·a·chance.¶
¶
Jeff·is·traveling·at·the·moment,·but·you·can·e-mail·me·or·Jeff·directly·at·¶
jeff_skilling@hotmail.com.¶
¶
Thanks·for·your·help.··SRS¶
¶
_____¶
Get·your·FREE·download·of·MSN·Explorer·at·http://explorer.msn.com/intl.asp¶

The lines marked in **red** are the metadata that is not included and those marked in **blue** are the earlier message content that shouldn't be included, as it is not part of the current message written by the current sender.
The CFL extraction program selects only the lines marked in **purple. T**hese are:

1. MessageId: used to provide the path for the email on its way out of Steven Kean's actual email address within Enron.
2. Date: The date and time of transmission of this message
3. From: Sender.
4. To: Addressee.
5. Subject: What it was about or related to.
6. Cc: who was copied in to the email (if anyone)
7. Bcc: who was blind copied into the email (if anyone).
8. The actual message.

In addition, it marks all lines except the actual message written by the sender with a delimiter so that concordance programs such as WordSmith can be instructed to ignore the text between such delimiters. The delimiters used in the program are < and >.

Programming procedure

Starting at the top of each file all lines up to and including the line that starts with **Subject:** are marked with the delimiters as shown below:
<Message-ID: 16170877.1075858881106.JavaMail.evans@thyme>>
<Date: Wed, 10 Oct 2001 11:49:31 -0700 (PDT)>
<From: j..kean@enron.com>
<To: jean.ryall@enron.com>
If the line starts with **To:** the program finds the line index position of the **@** sign and the position of the first full top before that sign. The text in between is the surname of the addressee, and the program stores this is a special variable for checking in stage 3(c) below.
If the line starts with Subject: the program adds delimiters
<Subject: FW: Wallis Jefferson>
Each delimited line is stored in a holding file in the computer memory, named HeaderData, awaiting final writing to the main file.

When it has identified the subject line the program proceeds using a different line testing function.
This function proceeds as follows:

1. If the line starts with Mime-, Content- or X- it ignores the line.
   (Lines starting with X- have been introduced by the CMU extraction program, and the Mime and Content lines are part of the formatting instructions for the email program used by the email custodian.)
2. If the line starts with **Cc:** or **Bcc:** the line is surrounded by the < and > delimiters and added to the holding file.
3. Otherwise the program tests for the following three primary indicators of the start of an earlier message to which the sender is responding:
   a. -----Original Message----- as in the example above. This normally starts either at the very start of a line or one character in.

b. Forwarded by indicating that any text following it was written by someone other than the sender.

for calendar. thanks df
---------- Forwarded by Drew Fossum/ET&S/Enron on 10/11/2000 08:57 AM ----------

c. The name of the sender, as in:

Thanks for the prompt info!! I hope we can help this guy get his power plant built. DF

From: Jeff Nielsen 10/02/2000 10:31 AM

The program uses the surname recovered from the **To:** line to check for the presence of the same surname on this line.

If any of these cases is found to be true the program stops reading the email, otherwise the line is added to a second holding file in memory, called BodyMessage. Mail that is simply forwarded will leave this line empty.
When the program stops reading, either because of one of the tests above or because it has reached the end of the current email message, it checks to see if the BodyMessage variable has anything in it. if it does it writes the full file path of the folder on a new line, the HeaderData on subsequent lines and finishes with the BodyMessage and a blank line. These are printed to a single file, so that all the messages are in a compact form in one location for review and analysis.
When all lines have been printed it proceeds to the next message in the selected folder for the current custodian.

The full final entry for the email on screen 4 is:

<Enron_maildir/kean-s/sent_items/122>
<Message-ID: 16170877.1075858881106.JavaMail.evans@thyme>>
<Date: Wed, 10 Oct 2001 11:49:31 -0700 (PDT)>
<From: j..kean@enron.com>
<To: jean.ryall@enron.com>
<Cc: sheri.sera@enron.com>
<Bcc: sheri.sera@enron.com>
What do you know about this guy?

The foregoing describes the current state of the program, and most of the functions described were in place from the start. However, during user viewing of the output, it became clear that more than one original email system was in use, and these were not consistent.
Three main changes were made:

- The original version of the program used the presence of a colon on a line to indicate the presence of a date or time entry, and this was taken as indicative of a an original message, where no other indication existed. As the instruction was to surround any such line with delimiters, this had the effect of removing a number of lines from the actual body, wherever the sender used a time reference to arrange an appointment. This condition needed to be removed.

- The conventions for the placing of the Original Message indicator were not consistent in all converted emails, so the program had to allow for change in this position, otherwise it did not recognise that the material following it was not in fact written by the sender, but the person to whom the sender was responding. While the **From:** and **To:** lines were correctly surrounded with delimiters, the body lines were not, so, unlike in case 1, too much text was being attributed to the sender.
- The complete absence of explicit indicators of an earlier message, other than the presence of the name of the person who had sent the original email to which the sender was responding, meant that again too much textual material was being attributed to current sender. This was corrected by the identification of the surname from the **To:** line, as described above.

David Woolls
13th May 2012

Appendix 2

University of Leeds' Arts and PVAC Faculty Research Ethics Committee's Light Touch Ethical Review decision.

Performance, Governance and Operations
Research & Innovation Service
Charles Thackrah Building
101 Clarendon Road
Leeds LS2 9LJ Tel: 0113 343 4873
Email: ResearchEthics@leeds.ac.uk

UNIVERSITY OF LEEDS

David Wright
School of English
University of Leeds
Leeds, LS2 9JT

Arts and PVAC Faculty Research Ethics Committee (PVAR)
University of Leeds

13 September 2014

Dear David

Title of study: **Stylistics versus statistics: A corpus linguistic approach to combining techniques in forensic authorship analysis using Enron emails.**
Ethics reference: **LTENG-001, Amendment August 2014**

I am pleased to inform you that the above application for light touch ethical review has been reviewed by a delegate of the Arts and PVAC (PVAR) Faculty Research Ethics Committee. I can confirm a favourable ethical opinion on the basis of the application form as of the date of this letter. The following documentation was considered:

Document	*Version*	*Date*
LTENG-001 email re revised research title.txt	1	28/08/14
LTENG-001 DavidWright (LTENG-001) Ethics Amendment_form.docx	1	20/05/13
LTENG-001 DW Light Touch Ethics Form Revised.docx	3	20/05/13
LTENG-001 EDRM Clean Up.pdf	1	20/05/13

Please notify the committee if you intend to make any further amendments to the original research as submitted at date of this approval as all changes must receive ethical approval prior to implementation. The amendment form is available at http://ris.leeds.ac.uk/EthicsAmendment.

Please note: You are expected to keep a record of all your approved documentation, as well as documents such as sample consent forms, and other documents relating to the study. This should be kept in your study file, which should be readily available for audit purposes. You will be given a two week notice period if your project is to be audited. There is a checklist listing examples of documents to be kept which is available at http://ris.leeds.ac.uk/EthicsAudits.

We welcome feedback on your experience of the ethical review process and suggestions for improvement. Please email any comments to ResearchEthics@leeds.ac.uk.

Yours sincerely

Jennifer Blaikie
Senior Research Ethics Administrator, Research & Innovation Service
On behalf of Dr William Rea, Chair, PVAR FREC

Appendix 3

Full breakdown of the eighty-author sample (EEC80)

Occupation	Author	Emails	Tokens	Occupation	Author	Emails	Tokens	Occupation	Author	Emails	Tokens
Presidents, CEOs and COOs	Beck, Sally	1,272	104,679	**Vice Presidents**	Arnold, John	1,038	26,283	**Managing Directors/ Directors**	Allen, Phillip	355	14,579
	Buy, Rick	481	14,817		Fossum, Drew	876	39,297		Blair, Lynn	804	22,040
	Delainey, David	677	26,778		Kean, Steven	1,118	33,751		Cash, Michelle	889	31,419
	Horton, Stanley	271	9,389		Presto, Kevin	816	23,077		Crandell, Sean	97	39,443
	Kitchen, Louise	911	25,899		Sager, Elizabeth	791	23,993		Dasovich, Jeff	2,855	170,316
	Lavorato, John	1,113	24,677		Sanders, Richard	959	17,466		Haedicke, Mark	800	18,450
	McConnell, Mike	613	34,188		Shackleton, Sara	3,465	148,518		StClair, Carol	990	55,147
	Shankman, Jeffrey	939	20,920		Steffes, Jim	1,201	35,284		Ward, Kim	509	20,262
	Skilling, Jeff	29	1,292		Taylor, Mark	1,548	71,849		Watson, Kimberly	797	23,007
	Whalley, Greg	139	3,861		Tycholiz, Barry	395	16,356		White, Stacey	370	12,384
		6,445	**266,500**			**12,207**	**435,874**			**8,466**	**407,047**
Lawyers	Bailey, Susan	11	605	**Managers**	Dorland, Chris	502	14,565	**Traders**	Bass, Eric	1,009	25,481
	Derrick, Jim	469	5,902		Farmer, Darren	772	24,502		Baughman, Don	39	3,602
	Hain, Mary	146	10,288		Forney, John	333	21,315		Germany, Chris	2,298	77,597
	Heard, Marie	613	24,269		Grigsby, Mike	478	19,277		Giron, Darron	452	15,415
	Hyvl, Dan	474	26,708		Kaminski, Vince	2,299	52,992		Guzman, Mark	246	13,227
	Jones, Tana	2,990	123,231		Keiser, Kam	327	13,108		Kuykendall, Tori	113	4,735
	Mann, Kay	3,055	103,512		Love, Phillip	981	36,471		Parks, Joe	213	4,498
	Nemec, Gerald	1,465	50,211		Richey, Cooper	137	4,833		Pereira, Susan	65	2,141
	Perlingiere, Debra	1,634	59,149		Ruscitti, Kevin	91	2,980		Scholtes, Diana	56	1,871
	Rapp, Bill	88	3,771		Schoolcraft, Darrell	264	7,475		Symes, Kate	1,227	58,754
		10,945	**407,646**			**6,184**	**197,518**			**5,718**	**207,321**
Analysts, Specialists, Associates	Campbell, Larry	226	28,482	**Assistants**	Elbertson, Janette	47	1,953				
	Causholli, Monika	181	7,161		Fleming, Rosalee	88	2,427				
	Lenhart, Matthew	1,059	21,064		Hillis, Kim	43	1,799				
	Quigley, Dutch	325	9,882		Lokay, Michelle	163	6,946				
	Rodrique, Robin	643	28,208		McVicker, Maureen	114	4,507				
	Rogers, Benjamin	616	20,552		Phillips, Cathy	32	1,938				
	Scott, Susan	880	61,482		Sera, Sherri	81	3,463				
	Smith, Matt	318	14,631		Sweet, Twanda	100	2,641				
	Weldon, Charles	231	9,720		Taylor, Liz	85	2,366				
	Williams, Bill	424	24,183		Thompson, Patti	54	2,894				
		4,903	**225,365**			**807**	**30,934**				