Identifying idiolect in forensic authorship attribution: an n-gram textbite approach

Alison Johnson & David Wright

University of Leeds

Abstract. *Forensic authorship attribution is concerned with identifying authors of disputed or anonymous documents, which are potentially evidential in legal cases, through the analysis of linguistic clues left behind by writers. The forensic linguist "approaches this problem of questioned authorship from the theoretical position that every native speaker has their own distinct and individual version of the language [. . .], their own idiolect" (Coulthard, 2004: 31). However, given the difficulty in empirically substantiating a theory of idiolect, there is growing concern in the field that it remains too abstract to be of practical use (Kredens, 2002; Grant, 2010; Turell, 2010). Stylistic, corpus, and computational approaches to text, however, are able to identify repeated collocational patterns, or n-grams, two to six word chunks of language, similar to the popular notion of soundbites: small segments of no more than a few seconds of speech that journalists are able to recognise as having news value and which characterise the important moments of talk. The soundbite offers an intriguing parallel for authorship attribution studies, with the following question arising: looking at any set of texts by any author, is it possible to identify 'n-gram textbites', small textual segments that characterise that author's writing, providing DNA-like chunks of identifying material? Drawing on a corpus of 63,000 emails and 2.5 million words written by 176 employees of the former American energy corporation Enron, a case study approach is adopted, first showing through stylistic analysis that one Enron employee repeatedly produces the same stylistic patterns of politely encoded directives in a way that may be considered habitual. Then a statistical experiment with the same case study author finds that word n-grams can assign anonymised email samples to him with success rates as high as 100%. This paper argues that, if sufficiently distinctive, these textbites are able to identify authors by reducing a mass of data to key segments that move us closer to the elusive concept of idiolect.*

Keywords: *Authorship attribution, email, Enron, idiolect, Jaccard, n-grams, style, textbites.*

Resumo. *A atribuição de autoria forense consiste em identificar os autores de documentos anónimos ou cuja autoria é contestada, e potencialmente elemento de prova em casos jurídicos, através da análise de pistas linguísticas deixadas pelos escritores. O linguista forense "aborda o problema da autoria questionada* *a partir do pressuposto teórico de que cada falante nativo de determinada língua possui a sua própria versão distinta e individualizada da língua [. . .], o seu próprio idiolecto" (Coulthard, 2004: 31). No entanto, considerando a dificuldade em sustentar empiricamente uma teoria do idiolecto, existe uma preocupação crescente nesta área relativamente ao facto de ser um conceito demasiado abstracto para ter utilidade prática (Kredens, 2002; Grant, 2010; Turell, 2010). As abordagens estilísticas, de corpora e computacionais ao texto, no entanto, permitem identificar padrões colocacionais repetidos, ou n-gramas, fragmentos linguísticos entre duas e seis palavras, semelhantes à conhecida noção de soundbites – pequenos segmentos de apenas alguns segundos de fala que os jornalistas conseguem identificar como possuindo valor noticioso, e que caracterizam momentos importantes da fala. O soundbite proporciona um fascinante paralelo para os estudos de atribuição de autoria, colocando-se a seguinte questão: observando um qualquer conjunto de textos de determinado autor, é possível identificar "textbites de n-gramas", pequenos segmentos de texto que caracterizam a escrita do autor, fornecendo segmentos de material identificativo semelhantes ao DNA? Partindo de um corpus de 63.000 emails e 2,5 milhões de palavras escritas por 176 funcionários da antiga empresa de energia americana Enron, realizamos um estudo de caso, que mostra, em primeiro lugar, recorrendo a uma análise estilística, que um funcionário da Enron produz repetidamente os mesmos padrões estilísticos de pedidos educadamente codificados, de uma forma que pode ser considerada habitual. De seguida, uma experiência estatística utilizando o mesmo autor do estudo de caso revela que os n-gramas de palavras permitem atribuir amostras de email anonimizadas a esse autor com taxas de sucesso da ordem dos 100%. Este artigo defende que, quando suficientemente distintivos, estes textbites têm capacidade para identificar os autores reduzindo um volume de dados massivo a segmentos-chave que nos aproximam do esquivo conceito de idiolecto.*

Palavras-chave: *Atribuição de autoria, email, Enron, idiolecto, Jaccard, n-gramas, estilo, textbites.*

Introduction

Journalists listening to live speech are able to single out soundbites, small segments of no more than a few seconds of speech that they recognise as having news value and which characterise the important moments of talk. Though 'soundbite syndrome' (Mazzoleni and Schulz, 1999: 251) is seen as a reductionist trend representing a "tendency towards shorter and more sensational texts or 'soundbite news"' (Knox, 2007: 28), it offers an intriguing parallel, thinking laterally, for authorship attribution studies. The following question arises: looking at any set of texts by any author, is it possible to identify 'n-gram textbites', small two to six word portions of text that characterise the writing of that author, providing DNA-like chunks of identifying material? Using the DNA metaphor, if textual chunks or 'textbites' (Ancu, 2011) are sufficiently distinctive, they may be able to identify authors by reducing the mass of words to meaningful and key segments that move us closer to the abstract and elusive concept of a 'linguistic fingerprint', which Coulthard (2004: 432) describes as "the linguistic 'impressions' created by a given speaker/writer [which] should be usable, just like a signature, to identify them". Coulthard (2004: 432) eschews the DNA metaphor as "unhelpful" and "impractical" because of the need for "massive databanks consisting of representative linguistic samples". The Enron email corpus (Cohen, 2009), however, consisting of a large population of authors (we use a 176-author corpus) and their 63,369 emails spanning several years (1998-2002), provides a real opportunity to search for distinctive and individuating small text segments that can be manually identified or computationally extracted for authorship attribution and author profiling in emails, adding to our knowledge of the generic lexical features of that text type and to writing characteristics of some of the employees who inhabit it. This "specialised corpus" (Flowerdew, 2004) therefore enables us to focus on identifying features which represent small sections of the individual genetic code, and allows us to evaluate their usefulness in authorship attribution against a background of a population-level reference corpus; thus, attempting to identify DNA-like textbites becomes a manageable undertaking.

Solan (2013) in his paper on intuition versus algorithm, details the current trends in forensic authorship analysis and the divergence between qualitative and stylistic approaches on the one hand and computational and statistical approaches on the other. Earlier, Solan and Tiersma (2004: 463) highlighted the importance for admissibility of expert evidence of methods which "employ linguistically motivated analyses in combination with quantitative tools", and the usefulness of corpora in combining such methods for the purposes of authorship analysis has been noted elsewhere (Cotterill, 2010; Kredens and Coulthard, 2012; Solan, 2013). There have been a number of recent studies which combine stylistic (Turell, 2010; Queralt and Turell, 2012), grammatical and multidimensional (Nini and Grant, 2013) approaches with quantitative methods. Turell (2010: 212), not only demonstrates the usefulness of a combined qualitative and quantitative approach in a real case where authorship is disputed, but also points out the benefits of empirical research for "expert witness performance", especially, noting Coulthard's (1994) contribution to the study of corpus-based evaluation of markers of authorship. Drawing on the benefits afforded to researchers by using corpora, and combining corpus analysis with stylistic and computational methods, we focus on word n-grams (groups of one, two, or more words) as features for analysis, arguing that these can be considered 'textbites' that are evidence of distinctive text-encoding by authors. This paper reports a case study of one Enron employee, James Derrick, in which both corpus stylistic and computational approaches are used to investigate and identify an individual's distinctive language use. This case study has three research aims:

1. Using a corpus linguistic, stylistic, and computational approach that employs both qualitative and quantitative methods, we investigate whether n-grams are distinctive and characteristic of an individual author's style, in this case their professional business email style.
2. Focusing on an individual's distinctive linguistic behaviour within a specific type of speech act—*please*-mitigated directives—we highlight the distinctive ways in which individual authors express this particular speech act, and how this speech act gives rise to author-distinct collocational preferences and identifiable word n-grams or n-gram textbites.
3. Using computational and statistical tests, we assess and evaluate the effectiveness of word n-grams of between one and six words in the successful attribution of large and small samples of emails to their actual author.

Using a combination of computational tools (*Wordsmith Tools*, (Scott, 2008); *Jangle*, (Woolls, 2013)) this case study utilises the Enron corpus in two ways; the first research aim relies on the corpus as a large-scale reference corpus representing a population of writers against which an individual's style can be compared, while the computational attribution experiment in the third research aim uses the Enron corpus as a large pool of 176 candidate authors. As a result of the triangulated approach, we are able to compare the results of the different methods to see whether they are complementary and confirmatory, to evaluate the accuracy and reliability of both approaches and underline the effectiveness of n-grams as important markers of authorship. Finally, we can consider the implications for authorship studies and methods for use in evidential cases.

N-grams, idiolect and email

The co-occurrence of lexical items within a short space of each other is known in corpus linguistics as 'collocation', and was first introduced by Firth (1957), later developed by Sinclair (1991), Stubbs (2001) and Hoey (2005) amongst others, and continues to be at the forefront of corpus linguistic research (Gries, 2013). The associations that people build between words and the ways in which they produce them in combinations is a psycholinguistic phenomenon, and has been analysed in terms of 'lexical phrases' (Nattinger and DeCarrico, 1992), 'formulaic sequences' (Wray, 2002, 2008), 'lexical priming' (Hoey, 2005; Pace-Sigge, 2013), and usage-based theories of lexico-grammar such as exemplar theory (Barlow, 2013). One factor which these different approaches have in common is that they all emphasise the personal or 'idiolectal' nature of preferences for certain word combinations and collocational patterns. Schmitt *et al.* (2004: 138), discussing formulaic sequences, argue that "it seems reasonable to assume that they [people] will also have their own unique store of formulaic sequences based on their own experience and language exposure". Similarly, Hoey (2005: 8–15), in his argument that collocation can only be accounted for if we assume that every word is primed for co-occurrence with other words, grammatical categories or pragmatic functions, claims 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)

Collocations and sequential 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 vocabulary and syntactic information in texts and an effective way of capturing words in context. Collocations and lexical strings in this study are referred to as n-grams and we also consider the extent to which they can be called 'n-gram textbites', small portions of text that characterise the writing of a particular author.

The idiolectal nature of collocations has been investigated, to a limited extent, in corpus linguistics. Mollin (2009) used a corpus-based statistical approach to analysing idiolectal collocations in the text of former UK Prime Minister, Tony Blair, and Barlow (2010, 2013) examined the relative frequencies of two- and three-word sequences used by five White House Press secretaries. Distinctive lexical choices and sequences have also been used as evidence of idiolect in authorship identification (Coulthard, 2004, 2013) and plagiarism detection (Johnson and Woolls, 2009; Culwin and Child, 2010) and the usefulness of formulaic sequences as style markers in forensic authorship attribution has been evaluated (Larner, 2014). In addition, word n-grams have been used as input features for automated techniques of distinguishing and identifying authors with both encouraging (Hoover, 2002, 2003; Coyotl-Morales *et al.*, 2006; Juola, 2013) and poor (Grieve, 2007; Sanderson and Guenter, 2006) results.

Frequencies of function words (Mosteller and Wallace, 1964; Burrows, 2002; Argamon and Levitan, 2005), and increasingly character n-grams (Chaski, 2007; Stamatatos, 2008, 2013; Luyckx and Daelemans, 2011; Koppel *et al.*, 2011, 2013) have been preferred over content words in authorship research, as it is argued that the latter are too heavily dependent on topic, context and writing situation (Stamatatos, 2009: 540; Koppel *et al.*, 2009: 11). However, as Coulthard (2013: 447–8) argues, although the occurrence of lexical items shared between topically related texts is significant in authorship attribution, "much more significant is the shared occurrence of co-selected items or what linguists call collocates". In other words, although different writers may write about the same topics or with the same purpose, the way in which they write about these things can quickly become linguistically unique. Moreover, computational authorship analysts (Argamon and Koppel, 2013: 300; Stamatatos, 2013: 428) are increasingly acknowledging that the feature sets used and results obtained from automated techniques are difficult, if not impossible, to explain and interpret in linguistic and stylistic terms. In contrast, building on the research of Nattinger and DeCarrico (1992), Wray (2002), and Hoey (2005) outlined above, theoretical explanations can be offered for variation observed between authors in their production of n-grams. They are identifiable and computationally accessible manifestations of individuals' idiolectal, primed and perhaps prefabricated collocational and phraseological preferences, resulting from a lifetime of unique linguistic exposure and experience. As such, word n-grams are linguistic features that may serve to bridge the gaps between cognitive, computational, and stylistic approaches to authorship analysis, an increasingly important aim of current research in the field (Nini and Grant, 2013; Argamon and Koppel, 2010, 2013).

The particular focus of this paper is on individuals' distinctive linguistic behaviour within a specific type of 'speech act' (Austin, 1962; Searle, 1969): *please*-mitigated directives. Linguistic research into email as a text type has continually identified the making of commands and requests for action as major communicative functions of emails (Sherblom, 1988; Baron, 1998; Gimenez, 2000). By extension, there has been a wave of research into politeness and requests in emails of various languages, particularly in institutional settings such as academia and commerce (Duthler, 2006; Lampert *et al.*, 2008; Merrison *et al.*, 2012; Chejnová, 2014). Variation has been a central component of much of this research, with comparisons being made in email behaviour across different contexts (organisational/educational) (Gains, 1999; Waldvogel, 2007), different languages, varieties and cultures (Lan, 2000; Bou-Franch and Lorenzo-Dus, 2008; Merrison *et al.*, 2012), genders (Van Den Eynden, 2012), within specific communities of practice (Luchjenbroers and Aldridge-Waddon, 2011) and different social and hierarchical participant relationships and roles (Bou-Franch, 2011; Economidou-Kogetsidis, 2011; Najeeb *et al.*, 2012). This paper aims to add to this list the idiolectal nature of directives, highlighting the distinctive ways in which individual authors express this particular speech act, and how this speech act gives rise to author-unique collocational preferences and identifiable n-grams or n-gram textbites. Since requests and orders are difficult to differentiate, given that both have the function of getting someone to do something, we follow Bax's (1986: 676) distinction, where in a request "the requesting person benefits from the future act" and there is a "reciprocal social relation" between the interactants, whereas in a directive "the person does not necessarily have to benefit from [the act]" and the addressee is in an "inferior social relation".

Data and method

This section introduces the Enron email corpus, the particular version we created as appropriate for authorship research, and the case study of one employee, James Derrick. We evaluate the combined case study, corpus comparison, and experimental evaluation approach and we introduce the statistical (Jaccard) and computational (*Jangle*) tools used and created for this research.

The Enron email corpus

The corpus used for this study is a dataset of 63,369 emails and 2,462,151 tokens written and sent between 1998 and 2002 by 176 employees of the former American energy company Enron. The Enron email data was first made publicly available online as part of the Federal Energy Regulatory Commission's legal investigation into the company's accounting malpractices (FERC, 2013), which led to the ultimate bankruptcy and demise of the company in the early 2000s. Many versions of the data have emerged across the web. The source of the database used in this study is the version collected and prepared by Carnegie Mellon University (CMU) (Cohen, 2009), as part of its 'Cognitive Assistant that Learns and Organises' (CALO) project. The data have subsequently been extracted, cleaned-up, and prepared specifically for the purposes of authorship attribution by Woolls (2012). The extraction process mined all of the sent emails from all of the various 'sent' folders for each of the authors, retaining only the newest email material in each thread, and removing any previous email conversation, to ensure that only the material written by the sender was included and available for analysis. This process was vital, to create a corpus suitable for authorship research, because, as it stands, the Enron corpus (Cohen, 2009) is unsuitable for this purpose. Each email in the corpus is accompanied by a range of metadata: the date and time the email was sent, along with the 'From:', 'To:', 'Subject:', 'Cc:' and 'Bcc:' fields, and the subject line (Example 1). In the cleaned-up corpus metadata is contained in angle brackets so that it is not considered by the computational tools used to analyse the authors' textual choices in this study.

```
(1)
<C:\EnronAuthorRef\derrick-j\sent\16>
<Message-ID: 764619.1075842926074.JavaMail.evans@thyme>
<Date: Fri, 1 Dec 2000 06:31:00 -0800 (PST)>
<From: james.derrick@enron.com>
<To: john.belew@enron.com>
<Subject: Re: Deferral Enrollment 2001>
John, I believe your message was sent to me by mistake.  I
am returning it to you.  Jim
```

The data were further cleaned by Wright, in particular where the authors' sent folders contained emails sent by their assistants or secretaries. Such emails were relatively easy to identify and remove and, 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, creating a separate set of files for this individual and treating them as an additional author. Finally, blank emails and emails containing only forwarded or copied and pasted material were also removed. The resulting 176 author, 63,369 email and 2,462,151 word corpus is a gold-standard corpus, making it particularly useful for authorship studies. It contains only naturally occurring email language data about which we can be sure of the 'executive author' (Love, 2002: 43), that is, the individual responsible for "formulating the expression of ideas and ma[king] word selections to produce the text" (Grant, 2008: 218). Since digital texts such as text messages, instant messages, and emails are becoming increasingly prominent in forensic casework, including email cases containing threatening, abusive, or defamatory material (e.g. Coulthard *et al.*, 2011: 538), this corpus represents a unique opportunity for empirical research with implications for authorship attribution casework.

James Derrick

The analyses focus on a case study of one Enron employee, James Derrick. Case studies are a beneficial method when "'how' or 'why' questions are being posed" and when particular "phenomena" are being studied (Yin, 2009: 2) and, in this case, we want to know how an individual makes directives. The case study approach is complemented by a corpus linguistic approach, which allows us to examine the uniqueness of his choices against the larger Enron population. Derrick was an in-house lawyer at Enron (Creamer *et al.*, 2009; Priebe *et al.*, 2005), in fact Enron's chief lawyer and General Counsel, or chief legal officer, with a staff of "200 in-house lawyers" and able to call on "more than 100 outside law firms from around the world" between 1991 and 2002 (Ahrens, 2006). He is a former Adjunct Professor of Law at the University of Texas Law School between 1984 and 1990, and is currently a managing partner in a US law firm.

He is represented in the corpus by 470 emails, a total of 5,902 tokens and 911 types. He was chosen as a case study because of the relatively small amount of data in his sent box. Although he has slightly more emails than the mean per author for the corpus (mean = 360), he has far fewer than the mean tokens per author (mean = 13,989). Derrick is therefore a curious case in that his emails are much shorter than average, but this is perhaps unsurprising given his status as chief lawyer and awareness of legal discovery[1]. He has a mean of 12.9 words per email, while the mean for the corpus is 41.74. Further, 155 of his 470 emails (32.9%) contain only one word. These single-word emails contain two types: *FYI (for your information* 22.5%) and *FYR (for your reference/review/records* 10.4%). These most minimal of messages are sent chiefly to one addressee for each type. 67% of the *FYI* emails are to j.harris@enron, one of the lawyers in Derrick's team, and all of the *FYR* single-word messages are sent to c.williams@enron. The relatively small amount of data available for Derrick, in theory, makes any analysis of style and any subsequent attribution tasks more difficult than it would be with an author with more data (Grant, 2007; Koppel *et al.*, 2013; Luyckx and Daelemans, 2011). The advantage of using a case study for authorship research is twofold. First, the small amount of data presents a similar challenge to that in real cases where data is often limited. And second, while a case study allows

[1] Legal *discovery* (called *disclosure* in the UK) requires the defendant (in a criminal case) or adverse party (in a civil case) to disclose anything that is asked for by the other side, which is needed in the preparation of the case prior to trial. This can include electronic documents such as email.

us to identify an individual authorial style with a high degree of detail, it also allows for comparison with the larger population of Enron employees. As Hoover (2010: 250) says, "style is [...] essentially, if not always explicitly, comparative. Any remark on a stylistic characteristic implies a comparison, even if it does not state one".

In the corpus stylistic analysis, Derrick's style is compared with that of the other 175 authors in the Enron corpus, which serves as a reference corpus against which the rarity or expectancy of particular n-grams are found in Derrick's emails. In the attribution experiment, the corpus represents a large pool of 176 candidate authors (including Derrick) from which to attempt to correctly attribute samples of Derrick's emails.

Computational tools

Two computational tools are used: *Wordsmith Tools* (Scott, 2008) and the specially-designed *Jangle* (Woolls, 2013). *Wordsmith Tools* is used to generate word lists and keyword lists from the data, as well as concordance results for any given search word or phrase. *CFL Jaccard N-gram Lexical Evaluator (Jangle)* is a Java-based program which is used to run the attribution experiments. The program automatically generates 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 program then runs a series of pair-wise comparisons, with the sample file being compared with one other set of emails at any one time (either the entire set of remaining emails of the author in question, or the entire email set of another Enron employee). Finally, *Jangle* then produces Jaccard results measuring how similar this pair of files is in terms of the word n-grams shared between them.

Jaccard's similarity coefficient

Jaccard's coefficient, (or 'Jaccard Index', 'Jaccard', or 'intersection distance') measures the fraction of the data that is shared between any two sets (A∩B) compared to all data available in the union of these two sets (A∪B) (Naumann and Herschel, 2010: 24). Jaccard is widely used as a similarity metric across a range of scientific disciplines such as ecology (Jaccard, 1912; Izsak and Price, 2001; Pottier *et al.*, 2013; Tang *et al.*, 2013) forensic psychology and crime linkage (Bennell and Jones, 2005; Woodhams *et al.*, 2008; Markson *et al.*, 2010) and document comparison (Rajaraman and Ullman, 2011; Deng *et al.*, 2012; Manasse, 2012). Drawing on these various different uses of the coefficient, Jaccard has been introduced into forensic authorship analysis as a way of measuring the similarity or distance between questioned and known documents based on a range of different linguistic features (Grant, 2010, 2013; Wright, 2012; Larner, 2014; Juola, 2013). In this study, Jaccard is used to measure the similarity between any two sets of emails based on the number of items—in this case word n-grams—found in both sets, divided by the number of total number of items in the two sets combined:

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

or

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

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. 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 this feature appears.

Experimental set up

The second part of the Derrick case study involves running an attribution experiment to evaluate how accurate and reliable n-grams are in successfully identifying Derrick as the author of a random sample of his emails. Ten different random samples of 20%, 15%, 10%, 5% and 2% of his emails were extracted from his 470 email set, resulting in a total of 50 test samples (Table 1).

Emails (n)	20% (93)	15% (70)	10% (46)	5% (23)	2% (9)
1	1,018	537	350	281	94
2	901	679	555	192	88
3	776	672	606	288	82
4	831	666	411	229	84
5	762	746	622	229	74
6	1,116	875	479	174	109
7	1,220	836	404	398	78
8	1,034	767	419	200	145
9	980	692	525	156	55
10	876	654	478	358	77
Mean	951	712	485	251	87
SD	141.7	92.4	86.4	79.8	24.2

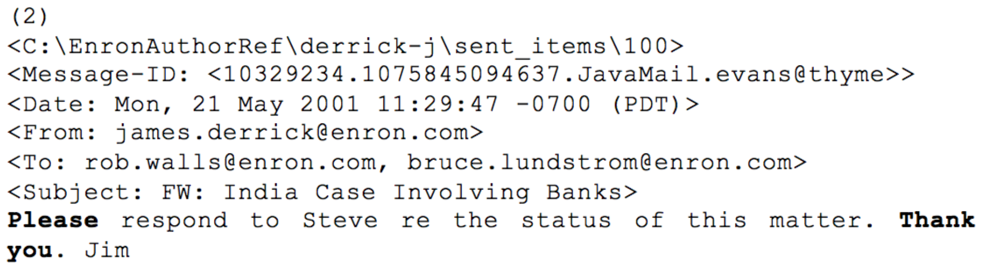
Table 1. Details of Derrick's sample sizes in terms of total tokens in the attribution task, including mean tokens and Standard Deviation (SD).

Previous stylometric studies have used test and sample datasets between 1,000 and 39,000 tokens in size (e.g. Hoover, 2004; Argamon and Levitan, 2005; Burrows, 2005; Labbé, 2007; Savoy, 2012). The sample sizes in this study range from the largest 20% samples at 1,220 and the smallest 2% sample at 55 tokens, and so are relatively small when compared with such stylometric research. They are more similar in size 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 341 tokens (Chaski, 2001; Grant, 2007; Rico-Sulayes, 2011).

Using *Jangle*, in each of the tests in this experiment, the extracted sample (2%, 5%, 10%, 15% or 20%) is compared against Derrick's remaining emails (either 98%, 95%, 90%, 85% or 80%) and the entire email sets of the other 175 Enron employees, in order to measure how similar the sample is to these other groups of texts. These comparisons are based on word n-grams ranging from single words through to six words in length. In a way which captures semantic as well as lexical and grammatical information, the n-grams used are sequential strings of words which co-occur within punctuated sentence boundaries, as opposed to a 'bag-of-words' approach, which does not take into account word order or punctuation. Whereas Derrick's samples are controlled for size, the comparison sets of the other Enron employees are not, and they range from as small as two emails and twenty tokens (Gretel Smith) to as large as 3,465 emails (Sara Shackleton) and 170,316 tokens (Jeff Dasovich). There is a total pool of 176 candidate authors used in this experiment (including Derrick), and this is relatively large in stylometric terms, with others using 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). There are exceptions such as Koppel *et al.* (2006, 2011), however, which use open candidate sets of thousands of potential authors. Overall though, the combination of small sample sizes and a large number of candidate authors makes the attribution task in this study a relatively difficult one.

Case study of James Derrick

This case study employs two different approaches to analysing James Derrick's use of n-grams. The first approach (in the subsection *Identifying Derrick's professional authorial style*) is a corpus stylistic one, which examines Derrick's professional authorial style, in particular through the distinctive ways in which he constructs directives (Example 2 below). He habitually uses email-initial please as well as a final thank you in appreciation of anticipated compliance, making these mitigated directives.

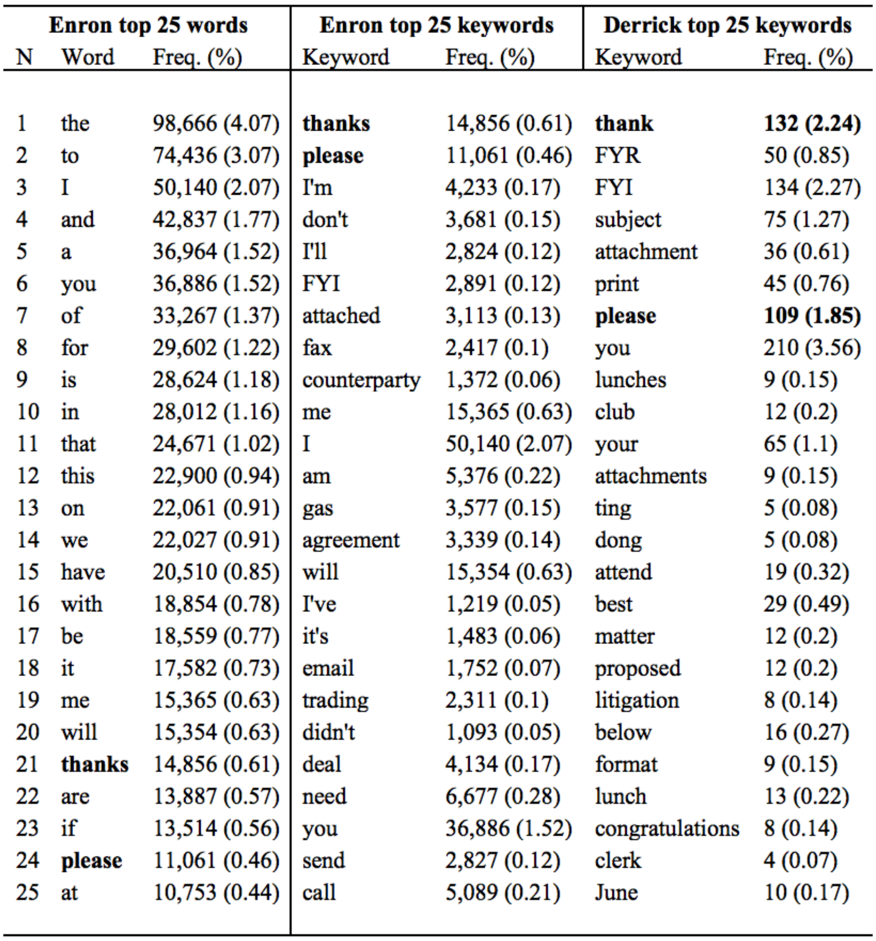


```
(2)
<C:\EnronAuthorRef\derrick-j\sent_items\100>
<Message-ID: <10329234.1075845094637.JavaMail.evans@thyme>>
<Date: Mon, 21 May 2001 11:29:47 -0700 (PDT)>
<From: james.derrick@enron.com>
<To: rob.walls@enron.com, bruce.lundstrom@enron.com>
<Subject: FW: India Case Involving Banks>
Please respond to Steve re the status of this matter. Thank
you. Jim
```

The second part of the case study (in the *Attribution task* subsection) uses a statistical and quantitative approach in the form of an attribution experiment using the Jaccard measure described above (in the *Experimental set up* section). Finally, the section on 'Derrick's discriminating n-grams' compares the results of these two approaches in terms of their reliability, complementarity, and compatibility.

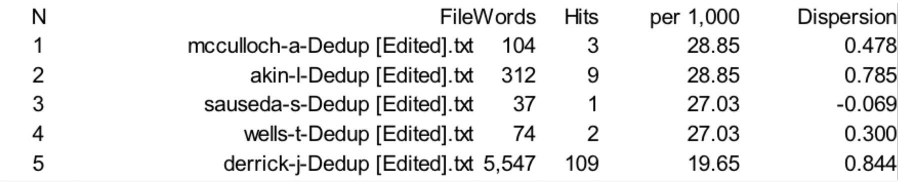
Identifying Derrick's professional authorial style

A starting point for any corpus analysis of authorial style is to create a frequency word list and then a keyword list (Hoover, 2009), to identify those words "whose frequency is unusually high [in a particular dataset] in comparison with some norm", because "keywords provide a useful way to characterise a text or genre" (Scott, 2010: 156). First, the Enron dataset is compared with the 450-million-word Corpus of Contemporary American English (COCA) (Davies, 2012), to identify which words are key in the Enron corpus. Second, Derrick's data are then compared with the whole Enron corpus, to identify which words are key in his emails, when tested against the Enron population from which he is drawn. *Wordsmith Tools* (Scott, 2008) was used to create a word list. In any wordlist, function words are the most common, and that is the case in the Enron corpus, as we see in Table 2, column 1 (*the*, *to*, *I*, *and*, *a*, *you*, *of*, *for*, *is*, and *in* make up the top 10). The most common lexical words, therefore, more usefully characterise a text than function words. In the Enron corpus the top two lexical words are *thanks* and *please* (Table 2, column 1), giving a clear indication that the Enron dataset is representative of a community of practice in which interlocutors are generally linguistically polite to one another, and also suggesting that requests or mitigated directives are this community's principal speech act. Indeed, *please* and *thanks* are found in 165 and 164 of the 176 employees' emails respectively. This is made even clearer by a keyword analysis, which finds that *thanks* and *please* are the top two key words in the entire Enron corpus (Table 2, column 2). In turn, a second keyword analysis comparing Derrick's emails with the emails of the other 175 authors (Table 2, column 3), finds that *thank (you)* is the most significant key word and *please* is the seventh most important in his emails. More importantly, in terms of the proportion of Derrick's vocabulary, *thank* accounts for 2.24% of his vocabulary, whereas it accounts for 0.05% of the Enron vocabulary, and please accounts for 1.85% against 0.46%. These two words, therefore, account for a total of 4.1% of Derrick's vocabulary, compared to 0.5% for the Enron authors in general, indicating that even within this corpus, which is generally indicative of very polite discourse, Derrick appears exceptionally linguistically polite.



Enron top 25 words			Enron top 25 keywords		Derrick top 25 keywords	
N	Word	Freq. (%)	Keyword	Freq. (%)	Keyword	Freq. (%)
1	the	98,666 (4.07)	**thanks**	14,856 (0.61)	**thank**	**132 (2.24)**
2	to	74,436 (3.07)	**please**	11,061 (0.46)	FYR	50 (0.85)
3	I	50,140 (2.07)	I'm	4,233 (0.17)	FYI	134 (2.27)
4	and	42,837 (1.77)	don't	3,681 (0.15)	subject	75 (1.27)
5	a	36,964 (1.52)	I'll	2,824 (0.12)	attachment	36 (0.61)
6	you	36,886 (1.52)	FYI	2,891 (0.12)	print	45 (0.76)
7	of	33,267 (1.37)	attached	3,113 (0.13)	**please**	**109 (1.85)**
8	for	29,602 (1.22)	fax	2,417 (0.1)	you	210 (3.56)
9	is	28,624 (1.18)	counterparty	1,372 (0.06)	lunches	9 (0.15)
10	in	28,012 (1.16)	me	15,365 (0.63)	club	12 (0.2)
11	that	24,671 (1.02)	I	50,140 (2.07)	your	65 (1.1)
12	this	22,900 (0.94)	am	5,376 (0.22)	attachments	9 (0.15)
13	on	22,061 (0.91)	gas	3,577 (0.15)	ting	5 (0.08)
14	we	22,027 (0.91)	agreement	3,339 (0.14)	dong	5 (0.08)
15	have	20,510 (0.85)	will	15,354 (0.63)	attend	19 (0.32)
16	with	18,854 (0.78)	I've	1,219 (0.05)	best	29 (0.49)
17	be	18,559 (0.77)	it's	1,483 (0.06)	matter	12 (0.2)
18	it	17,582 (0.73)	email	1,752 (0.07)	proposed	12 (0.2)
19	me	15,365 (0.63)	trading	2,311 (0.1)	litigation	8 (0.14)
20	will	15,354 (0.63)	didn't	1,093 (0.05)	below	16 (0.27)
21	**thanks**	14,856 (0.61)	deal	4,134 (0.17)	format	9 (0.15)
22	are	13,887 (0.57)	need	6,677 (0.28)	lunch	13 (0.22)
23	if	13,514 (0.56)	you	36,886 (1.52)	congratulations	8 (0.14)
24	**please**	11,061 (0.46)	send	2,827 (0.12)	clerk	4 (0.07)
25	at	10,753 (0.44)	call	5,089 (0.21)	June	10 (0.17)

Table 2. Enron top 25 words and keywords, compared with Derrick's top 25 key words.



N	File	Words	Hits	per 1,000	Dispersion
1	mcculloch-a-Dedup [Edited].txt	104	3	28.85	0.478
2	akin-l-Dedup [Edited].txt	312	9	28.85	0.785
3	sauseda-s-Dedup [Edited].txt	37	1	27.03	-0.069
4	wells-t-Dedup [Edited].txt	74	2	27.03	0.300
5	derrick-j-Dedup [Edited].txt	5,547	109	19.65	0.844

Figure 1. Rate per 1,000 words of please in the Enron corpus (top 5 users), along with dispersion.

This is further supported by Figure 1, which shows the top five users of *please* in the corpus. Derrick is the fifth most frequent user of *please* among the 176 employees, but looking at the dispersion rates in Figure 1 (the rate at which *please* is dispersed across the author's text) (Figure 1, column 6), and the number of words in each file (column 3), we can see that employees one to four do not really count, since their file size is well below 500 words and dispersion across their emails is generally low (apart from in the file for Akin: Akin-l-Dedup [Edited].txt). This makes Derrick arguably the most polite employee in the corpus, or at least the person who makes most politely mitigated directives.

Furthermore, the keyword analysis of the whole Enron corpus shows that *thanks* (top ranked keyword in Table 2, column 2) is far more popular than *thank* (ranked 444th keyword, with Derrick alone accounting for 10% of its frequency). In contrast, Derrick's usage is entirely the reverse of this; his first keyword is *thank* (which occurs with *you* in all 132 instances), while *thanks* is used only ten times and is a negative keyword in his dataset, meaning that he uses it significantly less than all of the other authors in the Enron corpus. To put this another way, in the Enron corpus *thanks* is more than 11 times more frequent than *thank you*, whereas for Derrick *thank you* is 15 times more frequent than *thanks*. Examples 3 and 4 are typical of the way that Derrick uses *please* and *thank you*, using pre-posed *please* to serve a politeness function and mitigate the face threat in the directive (Bax, 1986: 688). He follows up with *thank you*, assuming compliance, punctuated with a full stop and followed by his name. As such, not only is Derrick a linguistically polite communicant, but these quantitative and preliminary qualitative results indicate that he is polite in such a way that is very distinctive in this large population of Enron writers. In addition, his use of *Thank you*+full stop, to follow up his directive, indicates his status and authority to make these orders.

```
(3)
<C:\EnronAuthorRef\derrick-j\sent_items\405>
<From: james.derrick@enron.com>
<To: rob.walls@enron.com>
<Subject: Fw: Corporate Legal Allocations to EEL>
RW, please respond as you deem appropriate.  Thank you. Jim
```

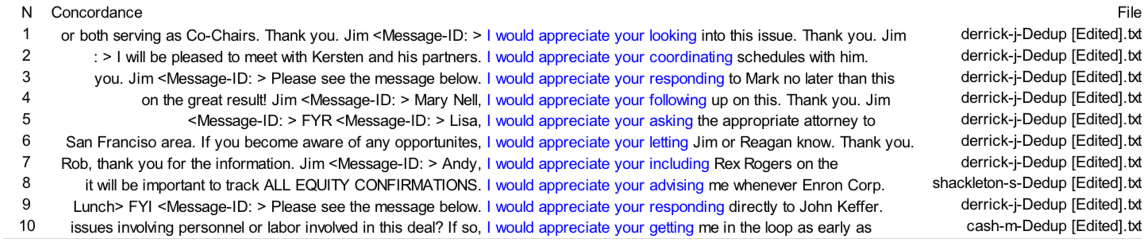
Given the obvious frequency, saliency and significance of linguistic politeness in the use of *please* and *thank(s)* in the Enron corpus, and in particular in Derrick's emails, we might say that this marks the construction of politeness and politeness strategies as an important aspect of his professional authorial style within the Enron corporation, a phenomenon which we deal with in detail below.

```
(4)
<C:\EnronAuthorRef\derrick-j\sent_items\376>
<From: james.derrick@enron.com>
<To:john.ale@enron.com,drew.fossum@enron.com,e..haedicke@enron
.com,>
<Subject: Jones, Day>
Andy Edison has notified me that Jones, Day is representing a
party in a litigation matter adverse to Enron. Please notify
me if you are aware of any work that Jones, Day is performing
for any Enron entity or for any lender or underwriter on an
Enron matter.  Thank you. Jim
```

Please and other politeness phenomena

The most effective way of analysing Derrick's distinctive construction of politely mitigated directives is through taking a collocational approach. *Wordsmith Tools* was used to identify the collocates that Derrick co-selects with *please* (Figure 2). The first point to note is that of Derrick's 109 *please* occurrences, 69 (63%) appear in message-initial position (after the greeting or no greeting). This is a marked use of *please* in relation to the Enron corpus; of the 10,952 instances of *please* that are found in the emails of the other 175 authors, only 2,092 (19%) are message-initial across 118 of the authors (e.g. Example 5), though this pattern is apparent in Table 3.

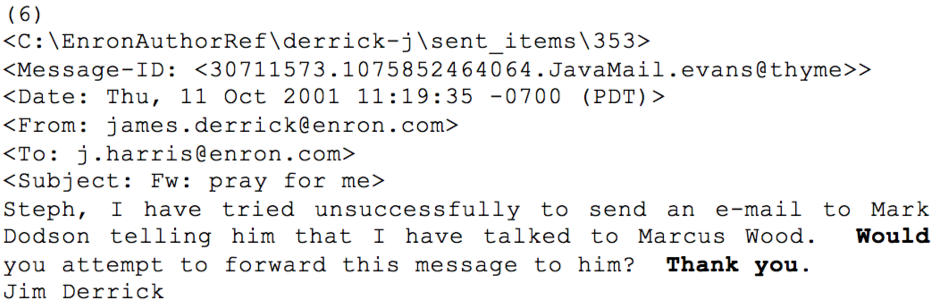
```
(5)
<C:\EnronAuthorRef\dasovich-j\sent\4502>
<Message-ID: <16191888.1075843889453.JavaMail.evans@thyme>>
<Date: Thu, 10 May 2001 06:20:00 -0700 (PDT)>
<From: jeff.dasovich@enron.com>
<To: joseph.alamo@enron.com>
<Subject:>
Hi:
please add $4 of tolls for each day that I went to Sac this
week.
Thanks,
Jeff
```



N	Concordance	File
1	or both serving as Co-Chairs. Thank you. Jim <Message-ID: > I would appreciate your looking into this issue. Thank you. Jim	derrick-j-Dedup [Edited].txt
2	: > I will be pleased to meet with Kersten and his partners. I would appreciate your coordinating schedules with him.	derrick-j-Dedup [Edited].txt
3	you. Jim <Message-ID: > Please see the message below. I would appreciate your responding to Mark no later than this	derrick-j-Dedup [Edited].txt
4	on the great result! Jim <Message-ID: > Mary Nell, I would appreciate your following up on this. Thank you. Jim	derrick-j-Dedup [Edited].txt
5	<Message-ID: > FYR <Message-ID: > Lisa, I would appreciate your asking the appropriate attorney to	derrick-j-Dedup [Edited].txt
6	San Franciso area. If you become aware of any opportunites, I would appreciate your letting Jim or Reagan know. Thank you.	derrick-j-Dedup [Edited].txt
7	Rob, thank you for the information. Jim <Message-ID: > Andy, I would appreciate your including Rex Rogers on the	derrick-j-Dedup [Edited].txt
8	it will be important to track ALL EQUITY CONFIRMATIONS. I would appreciate your advising me whenever Enron Corp.	shackleton-s-Dedup [Edited].txt
9	Lunch> FYI <Message-ID: > Please see the message below. I would appreciate your responding directly to John Keffer.	derrick-j-Dedup [Edited].txt
10	issues involving personnel or labor involved in this deal? If so, I would appreciate your getting me in the loop as early as	cash-m-Dedup [Edited].txt

Figure 2. Concordance of *I would appreciate your* verb-*ing* in the Enron corpus.

Instead, as well as using *please* within the body of the email, the other authors in the Enron corpus often modalise the directive with *can*, *could*, *would*, or *will* (See the L2 collocates in Table 3 – *can you please call*; *could you please send*; *will you please review*.). Derrick never uses *can*, *could*, or *will* to modalise directives, though he once uses *would* (Example 6). In this case, though, he grammatically marks the sentence as a question, signalling his lack of confidence that it can be carried out. Therefore, the use of unmodalised *please* directives and the use of *please* in email-initial position can be considered highly distinctive of Derrick. When message-initial *please* is combined with his choice of greeting form, Derrick's style stands out; he never uses *Hi* or *Hi:* (as in Dasovich, Example 5) and he consistently uses comma after the addressee's name, tending to follow up with *thank you* (Example 6). *Thank you* is discussed further below.



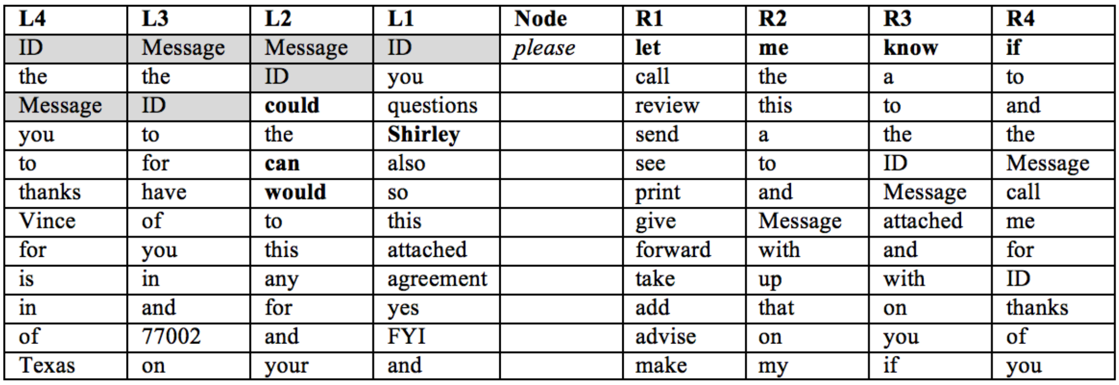
```
(6)
<C:\EnronAuthorRef\derrick-j\sent_items\353>
<Message-ID: <30711573.1075852464064.JavaMail.evans@thyme>>
<Date: Thu, 11 Oct 2001 11:19:35 -0700 (PDT)>
<From: james.derrick@enron.com>
<To: j.harris@enron.com>
<Subject: Fw: pray for me>
Steph, I have tried unsuccessfully to send an e-mail to Mark
Dodson telling him that I have talked to Marcus Wood.  Would
you attempt to forward this message to him?  Thank you.
Jim Derrick
```

On eight occasions he constructs indirect directives with *would* (Figure 2), in all cases with *your*+present participle, using the subjunctive mood with a gerund, producing the collocational string: *I would appreciate your* verb-*ing*. This grammatical choice is extremely rare in the wider corpus, with only three other examples, one each for Cash and Shackleton (lines 8 and 10 in Figure 2) and one for Beck (*I would appreciate your not sharing this plan*). The highly distinctive indirect directive *I would appreciate your* verb+*ing* is an important variation from the *please*+directive form, but, significantly, it is not a choice for 172 of the Enron authors. The Enron corpus shows the other authors' grammatical choices with *I would appreciate* as containing the following six patterns and choices (in decreasing order of frequency):

1. *I would appreciate* +NP (e.g. *a quick call*; *discretion*; *it.*) 56 occurrences
2. *I would appreciate your* +noun (e.g. *I would appreciate your assistance*) 24
3. *I would appreciate it if you could* +verb (e.g. *I would appreciate it if you could join us*) 10
4. *I would appreciate* +verb-*ing* (e.g. *I would appreciate hearing about*) 8
5. *I would appreciate if you could* +verb (e.g. *I would appreciate if you could call him*) 7
6. *I would appreciate you* +verb-*ing* (e.g. *I would appreciate you coordinating*) 2

In terms of similar grammatical constructions to Derrick's subjunctive + gerund, we can see that the most frequent is to nominalise (pattern 1 or 2): *I would appreciate a quick call*, or *I would appreciate your assistance/feedback/help/input/views*, rather than *I would appreciate your calling/assisting*. After that patterns 3 and 5 use the conditional *if* and patterns 4 and 6 use the present participle with or without *you*, but not *your*. The corpus norms therefore seem to be a grammatical choice of nominalising, using the conditional, or using the present participle, with Derrick's pattern being almost unique to him and shared with only three other authors (Shackleton, Cash and Beck). However, these authors, do not use the subjunctive plus gerund pattern exclusively with *I would appreciate*, as Derrick does. Instead they vary across the other six patterns (Shackleton uses 1, 2, and 5; Cash uses 1, 2, 3, and 4; Beck uses 1 and 2). The exclusive use of the subjunctive plus gerund is therefore a unique marker of Derrick's style, and importantly, as Turell (2010:

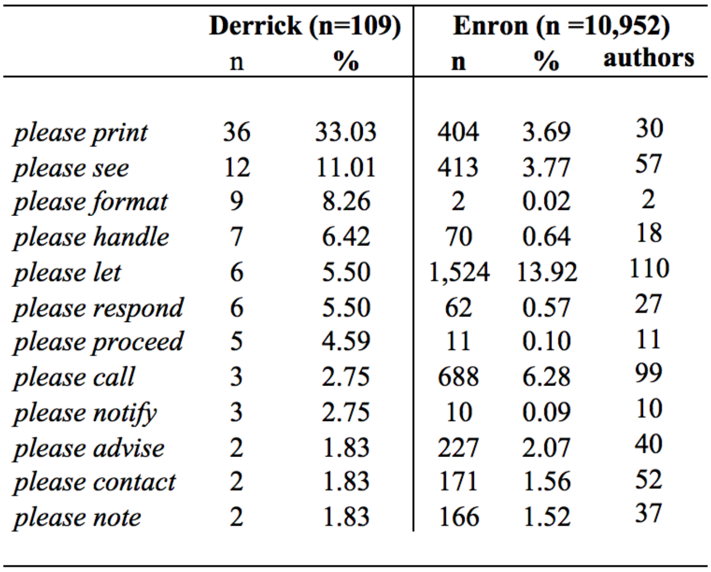
213) notes in relation to other grammatical markers, it is a grammatical marker that is "sociolinguistically constrained". She notes that "there are authorship markers, which apart from carrying grammatical substance, may contain sociolinguistic information about the author" (Turell, 2010: 214–215). In Derrick's case it identifies him as the most grammatically distinct (some might say most grammatically correct) user of this string. *I would appreciate your* verb-*ing* becomes a 5-gram textbite for him. We return to this below.



L4	L3	L2	L1	Node	R1	R2	R3	R4
ID	Message	Message	ID	*please*	**let**	**me**	**know**	**if**
the	the	ID	you		call	the	a	to
Message	ID	**could**	questions		review	this	to	and
you	to	the	**Shirley**		send	a	the	the
to	for	**can**	also		see	to	ID	Message
thanks	have	**would**	so		print	and	Message	call
Vince	of	to	this		give	Message	attached	me
for	you	this	attached		forward	with	and	for
is	in	any	agreement		take	up	with	ID
in	and	for	yes		add	that	on	thanks
of	77002	and	FYI		advise	on	you	of
Texas	on	your	and		make	my	if	you

Table 3. Enron authors' collocation patterns to the left (L1 to L4) and right (R1 to R4) of the node: *please* (Greyed out cells point to the message-initial placement of *please*).

It would not be possible to deal with please-mitigated directives in email without referring to the most frequent string of collocates with *please*: *please let me know* (Table 3). The top line of the R1 to R4 collocates in Table 3 shows this string; in the Enron corpus 110 of 176 authors use this, including Derrick. Taking the top verbs to the right of please, the R1 collocates show those which are most frequent for the Enron authors: *let*, *call*, *review*, *send*. Comparing the Enron top 10 lexical verb collocates with Derrick's in Table 4 (*print*, *see*, *format*, *handle*, *let*, *respond*, *proceed*, *call*, *notify*, *advise*), we find that only *let*, *see* and *print* are in the top 10 for the other Enron authors and we also see that *print* and *see* are Derrick's most frequent lexical verbs, rather than *let* and *call* for the other Enron employees, pointing to Derrick's reduced use of this phrase. *Please let me know* is found in Derrick's emails three times less frequently than it is across the corpus generally. At this point it is worth raising a problem with this use of the reference population to make comparisons. While *please let me know*, is three times more frequent in the wider Enron corpus than in Derrick's emails, Derrick is not the least common user. Taking the normalised frequency of rate per thousand words (using *Wordsmith*), Derrick appears 67th out of 110 in terms of frequency, so there are 43 authors who use it less than him and 66 who do not use it at all. In terms of dispersion (the extent to which the phrase is distributed across all his emails) he is 60th out of 110 authors who use it. So while comparing Derrick's use with the reference corpus as a whole shows his use to be distinctive, when we compare him with the individual authors in the Enron corpus, the population shows a wide variation in use from 7.30 uses per thousand words for Shapiro to 0.03 for Kean, with Derrick using it 0.54 times per 1,000 words. Nevertheless Derrick's twelve verb collocates account for 93 (85.3%) instances of all his *please* occurrences and the fact that many of Derrick's recurrent collocates are far more common in his emails than in the Enron corpus generally (Table 4), indicates that these are likely to be important bigrams in the experimental approach (*Attribution task* section). When looking at the percentage use of these collocates, Derrick uses *please see* three times more often than it appears in the emails of the other 175 authors (11.01/3.77), *please print* nine times more often, *please handle* and *please respond* ten times more, *please notify* 30 times more, *please proceed* 46 times more and *please format* as much as 413 times more often. Given that the raw frequencies are low in some cases (e.g. *please format*: 9 to 2 occurrences), some of the collocates are evidence of more consistent collocational differences, and, as we note above, the comparison with the Enron reference corpus masks some of the individual differences. Most of these collocations are used by a good number of authors, from 10 with *please notify* to 57 with *please see.*



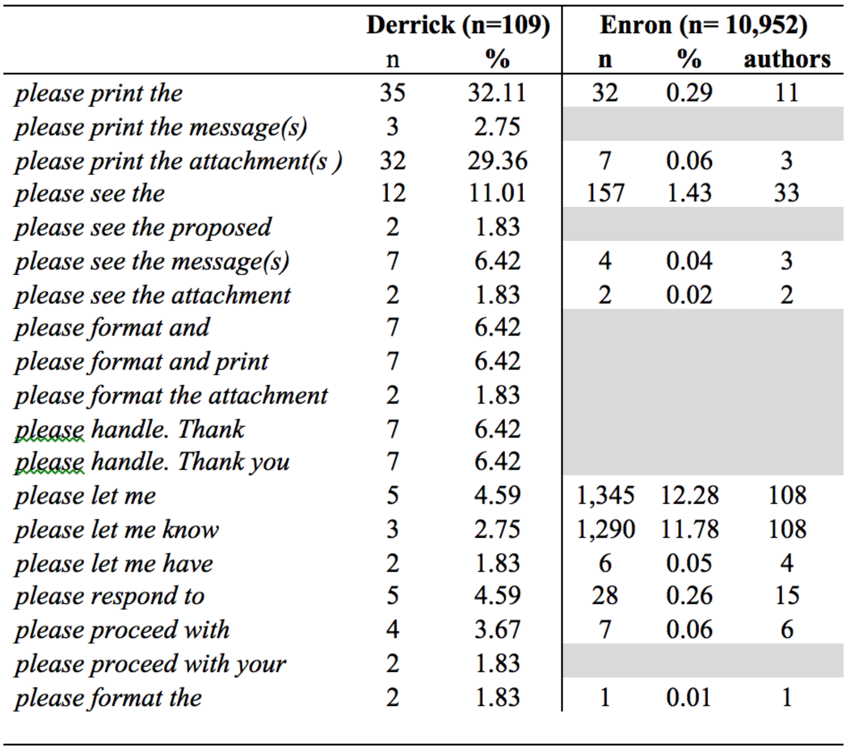
	Derrick (n=109)		**Enron (n =10,952)**		
	n	**%**	**n**	**%**	**authors**
please print	36	33.03	404	3.69	30
please see	12	11.01	413	3.77	57
please format	9	8.26	2	0.02	2
please handle	7	6.42	70	0.64	18
please let	6	5.50	1,524	13.92	110
please respond	6	5.50	62	0.57	27
please proceed	5	4.59	11	0.10	11
please call	3	2.75	688	6.28	99
please notify	3	2.75	10	0.09	10
please advise	2	1.83	227	2.07	40
please contact	2	1.83	171	1.56	52
please note	2	1.83	166	1.52	37

Table 4. Derrick's lexical verb collocates of *please*.

As noted above, *please format* is not found frequently elsewhere and *please handle* is a particularly important bigram (two-word n-gram) for Derrick. All of the seven instances of *please handle* in Derrick's emails are message initial and intransitive (Example 8). Of the 70 instances in the rest of the Enron corpus, used by 18 authors, 12 are transitive (e.g. *TK, Would you please handle this. Thanks, Kim*) and a further 32 are non-message initial (e.g. *Ray and Rob: Can you 3 help her out on 1 & 2? Tracy: Please handle*.) This leaves 29 instances of *please handle* in the Enron corpus that are intransitive and message initial, and these instances are shared by eight of the other 175 authors. However, all of Derrick's instances of *please handle* are consistently followed by *Thank you* and a sign-off using his name (Example 8). None of the 29 instances elsewhere in the Enron corpus shares this pattern; instead, they co-occur with the far more common *thanks*, just the author's name, *for me*, or by nothing at all (Examples 9-12). As such, message-initial, intransitive *please handle* followed by *thank you* and the author's name is only used by Derrick in the Enron corpus, and so is entirely unique and individuating of his email style.

(8) Please handle. Thank you. Jim
(9) Please handle. Thanks [email by Kevin Presto]
(10) Please handle. Mark [email by Mark Haedicke]
(11) Please handle. [email by Richard Sanders]
(12) please handle for me. thanks. mhc [by Michelle Cash]

Table 5 shows the recurring three and four word n-grams with *please* in Derrick's emails, as a proportion of all 109 instances of *please* in his dataset. The greyed out cells in the Enron column show that, in addition to '*please handle. Thank you.*', six additional tri- and four-grams are unique to Derrick: *please print the message(s), please see the proposed, please format and (print), please format the attachment*, and *please proceed with your.* All the rest are distinctive of him, apart from *please let me* and *please let me know* (discussed above), having much higher frequencies for Derrick that the reference corpus generally. To illustrate Derrick's distinctiveness, for example, in the case of *please print the*, which he uses more than 100 times more than the other authors (32.11% versus 0.29%), other authors use *this*, *and*, and *attachment* more as collocates of *please print.*



	Derrick (n=109)		**Enron (n= 10,952)**		
	n	**%**	**n**	**%**	**authors**
please print the	35	32.11	32	0.29	11
please print the message(s)	3	2.75			
please print the attachment(s)	32	29.36	7	0.06	3
please see the	12	11.01	157	1.43	33
please see the proposed	2	1.83			
please see the message(s)	7	6.42	4	0.04	3
please see the attachment	2	1.83	2	0.02	2
please format and	7	6.42			
please format and print	7	6.42			
please format the attachment	2	1.83			
please handle. Thank	7	6.42			
please handle. Thank you	7	6.42			
please let me	5	4.59	1,345	12.28	108
please let me know	3	2.75	1,290	11.78	108
please let me have	2	1.83	6	0.05	4
please respond to	5	4.59	28	0.26	15
please proceed with	4	3.67	7	0.06	6
please proceed with your	2	1.83			
please format the	2	1.83	1	0.01	1

Table 5. Derrick's recurring three and four word n-grams starting with *please*.

Although the stylistic analysis has clearly highlighted distinctive patterns of encoding polite directives in Derrick's emails, a further examination is even more revealing. For example, in 29 of the 32 instances of *please print the attachment(s)*, he is entirely consistent in following this with *thank you* (Example 13). In contrast, none of the seven instances in the rest of the Enron data is followed by *thank you* or indeed any sign-off at all, with *for me* and *thanks* being the most recurrent patterns (Examples 14 and 15). As such, as with *please handle* above, *please print the attachment(s)* + *thank you* is a pattern unique to Derrick in this corpus.

(13) Please print the attachment. Thank you.
(14) Please print the attachment for me in color
(15) Please print the attachments for me. Thanks,

Similarly, all seven of his uses of *please see the message(s)* are followed by *below* (Examples 16 and 17). However, although this four-gram appears four times in the remaining Enron data, only three of them are followed by *below*, each in a different author (Example 18).

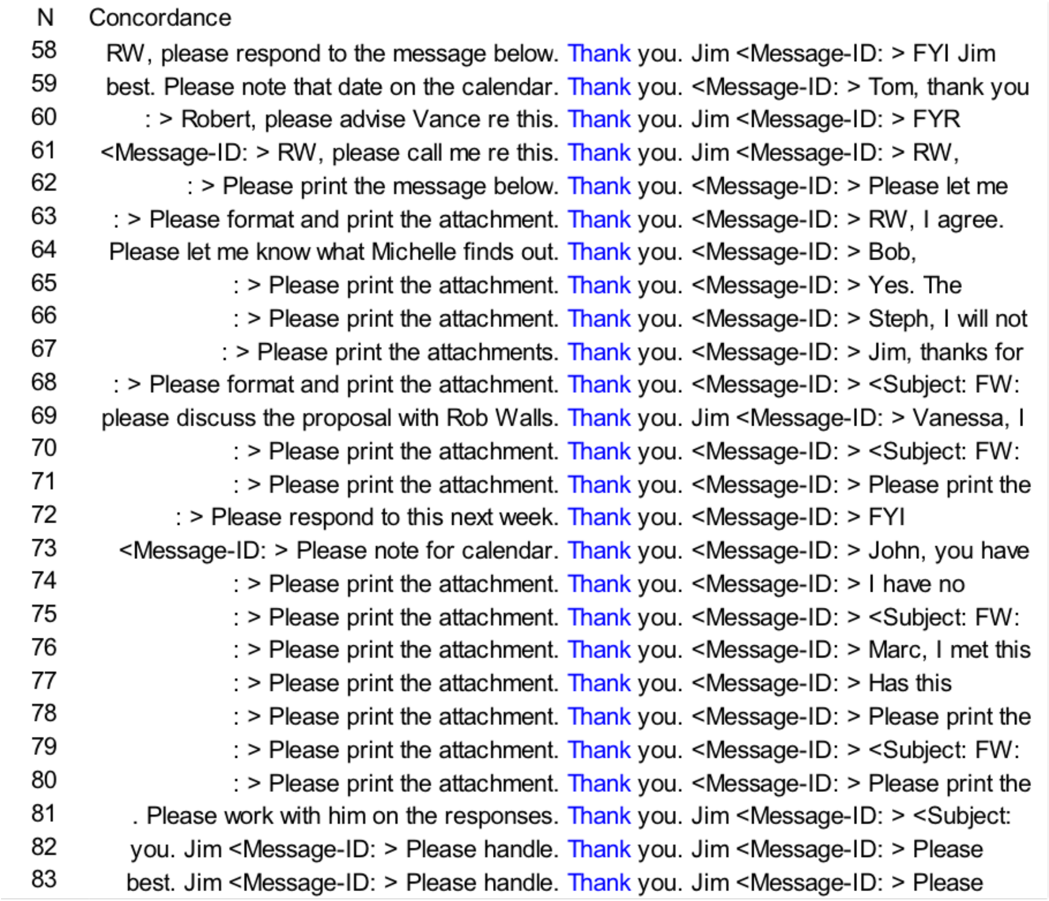
```
(16) Lisa, please see the message below from Gardere. Thank
you. Jim
(17) Please see the message below. I would appreciate [...]
(18) Kristina, please see the messages below and let me [...]
```

Furthermore, as Grant and Baker (2001) and others have explained, when style markers are aggregated, they quickly make a style unique. Taking together *please respond to*, *please print the attachment*, and *please see the* (from Table 5), and *I would appreciate your* (discussed above), Derrick is the only author out of all 176 to use these four n-grams, making these highly identifiable textbites for Derrick in relation to his use of mitigated directives.

Finally, apart from the linguistic politeness of Derrick's *please*-mitigated directives, he is additionally polite in thanking his addressees. He follows up with *Thank you*+full stop (See Figure 3.). This co-selection is very frequent in Derrick's dataset; of his 109 pleases, 93 (85.3%) co-occur with *thank you*. *Thank you*, co-occurring with *please*, is therefore distinctive of Derrick's emails. However, in the wider Enron corpus, although far less common than *thanks* (14,856) *thank you* is still used 1,144 times by 125 of the other 175 authors. However, the way in which Derrick uses *thank you* is still distinctive, as he consistently uses it as a sign-off (Wright, 2013 also found that Enron sign-offs can be remarkably distinctive.). Of the 132 instances of *thank you* in his emails, 109 (82.6%) are used as part of the email farewell or sign-off formula, either as a standalone *thank you*, or followed by his name, as seen in Figure 3. In the rest of the Enron corpus, 539 (47.1%) of the 1,114 instances of *thank you* are used either as standalone farewells or followed by the author's name, far fewer than in Derrick's data. The predominant pattern in Derrick's emails is to use *thank you* followed by only his first name (e.g. lines 81-83 in Figure 3), and he does this 40 times in 132 instances of *thank you* (30.3%), compared with only 172 out of 1,114 occurrences of *thank you* in the rest of the Enron corpus (15.4%). Overall, not only is Derrick's preference for *thank you* over *thanks* distinctive when compared with the other Enron authors, but his use of it is distinctive when compared with how those other 125 authors use it, most notably the ways in which he uses it as part of a sign-off, in particular as a standalone sign-off in combination with his first name only, and, most distinctively, as a follow-up to a mitigated directive.

Summary of findings

Overall, this stylistic analysis has focused on one particularly important aspect of Derrick's professional authorial style, the distinctive ways in which he lexico-grammatically frames politeness in directives mitigated with *please*. In the first instance, proportional frequency results show that the frequency with which Derrick uses *please* and *thank you* is marked when compared with the rest of the Enron data. Moreover, qualitative examination of the position in which Derrick uses these words, in terms of message-initial or final, and the combination of them, has revealed further distinctive patterns. Most importantly, though, a collocational n-gram approach has identified the ways in which lexical strings beginning with *please* can quickly become distinctive of Derrick, and several strings are unique. By combining as few as 4 distinctive strings, though individually these are shared across authors in the corpus, this combination becomes unique to Derrick. Even within a corpus in which politeness is a very frequent and salient linguistic strategy and directives are a common speech act, n-gram textbites can be found for an author. This can be ex-



N	Concordance
58	RW, please respond to the message below. Thank you. Jim <Message-ID: > FYI Jim
59	best. Please note that date on the calendar. Thank you. <Message-ID: > Tom, thank you
60	: > Robert, please advise Vance re this. Thank you. Jim <Message-ID: > FYR
61	<Message-ID: > RW, please call me re this. Thank you. Jim <Message-ID: > RW,
62	: > Please print the message below. Thank you. <Message-ID: > Please let me
63	: > Please format and print the attachment. Thank you. <Message-ID: > RW, I agree.
64	Please let me know what Michelle finds out. Thank you. <Message-ID: > Bob,
65	: > Please print the attachment. Thank you. <Message-ID: > Yes. The
66	: > Please print the attachment. Thank you. <Message-ID: > Steph, I will not
67	: > Please print the attachments. Thank you. <Message-ID: > Jim, thanks for
68	: > Please format and print the attachment. Thank you. <Message-ID: > <Subject: FW:
69	please discuss the proposal with Rob Walls. Thank you. Jim <Message-ID: > Vanessa, I
70	: > Please print the attachment. Thank you. <Message-ID: > <Subject: FW:
71	: > Please print the attachment. Thank you. <Message-ID: > Please print the
72	: > Please respond to this next week. Thank you. <Message-ID: > FYI
73	<Message-ID: > Please note for calendar. Thank you. <Message-ID: > John, you have
74	: > Please print the attachment. Thank you. <Message-ID: > I have no
75	: > Please print the attachment. Thank you. <Message-ID: > <Subject: FW:
76	: > Please print the attachment. Thank you. <Message-ID: > Marc, I met this
77	: > Please print the attachment. Thank you. <Message-ID: > Has this
78	: > Please print the attachment. Thank you. <Message-ID: > Please print the
79	: > Please print the attachment. Thank you. <Message-ID: > <Subject: FW:
80	: > Please print the attachment. Thank you. <Message-ID: > Please print the
81	. Please work with him on the responses. Thank you. Jim <Message-ID: > <Subject:
82	you. Jim <Message-ID: > Please handle. Thank you. Jim <Message-ID: > Please
83	best. Jim <Message-ID: > Please handle. Thank you. Jim <Message-ID: > Please

Figure 3. Screenshot of *Wordsmith Tools* concordances showing Derrick's use of *Thank you*.

plained with reference to the relationship between cognition and linguistic output central to lexical priming and formulaic sequences (Hoey, 2005; Wray, 2002). It may be that the repetitive nature with which Derrick constructs these sequences has 'primed' and stored these collocation patterns in his mind, and he reproduces these in distinctive ways. Or it may be that because he repeatedly finds himself within the relevant communicative context with regard to making the same directives, with the same purpose and recipient(s), the expression of these directives has become formulaic.

Beyond this speculation, this section has shown that as the number of n-grams in the textbite increases, so too does both the rarity and the distinctiveness of the lexical strings in question. This aligns with the findings of others (Coulthard, 2004; Johnson and Woolls, 2009; Culwin and Child, 2010). However, there are a number of ways in which this particular set of results for Derrick is exceptionally noteworthy. First, all of the n-grams presented here have been used at least twice by Derrick in his emails. Thus, as well as displaying inter-author differences in the corpus, and being distinctive of or even unique to his style, there is also a degree of intra-author consistency, rather than being strings used only once, which we might expect to be rarer generally, when tested against a population. Second, many of the n-grams identified as being distinctive of Derrick refer to things to do with emails and emailing; words such as *print*, *format*, *attachment*, and *message* are all related to computer-mediated communication and emails in particular.

The significance of this is that although all 176 authors in the Enron corpus share the same email mode of communication, the explicit references to such shared features that Derrick makes are framed in such a way that is distinctive of his style. However, most importantly, the Enron corpus is a very polite one; *please* was the second most common keyword in the corpus (Table 2), being used 11,061 times by 165 of the 176 employees. Despite this, the collocational analysis has found politeness structures that are particularly distinctive of Derrick's style, such as the indirect directive with the gerund (*I would appreciate your* verb+*ing*), when compared against this very polite corpus, a population of writers whom one may assume would be most stylistically similar to him.

Moreover, Derrick's choices become individuating very quickly; even some bigrams are distinctive of his emails, and, by the time the lexical string is three or four words long, most of the patterns are unique to him. This is considerably shorter than the required ten word strings postulated by Coulthard (2004) to be unique, and even shorter than the six reported by Culwin and Child (2010). Although these two studies used the web as corpus—a far larger comparison corpus—it can be argued that the specificity and relevance of the Enron corpus used here makes the comparison equally valid. Regardless, using a stylistic approach, and using the Enron corpus as a reference dataset, these results provide evidence to suggest that bigrams, trigrams, and four-grams have strong discriminatory potential for authorship analysis, particularly when focused on a particular function of language use, in this case the polite encoding of directives. The implication of this is that n-grams may offer a powerful means of attributing the authorship of disputed texts, even when applied in a pool of candidate writers writing within the same mode of communication and within the same community of practice. The remainder of the case study which follows sets out to test this hypothesis in an authorship attribution task.

Attribution task

This section reports the results of the attribution experiment outlined in the *Experimental set up* section. Ten random samples of 20%, 15%, 10%, 5% and 2% of Derrick's emails—ranging between 1,220 and 55 tokens in size—were extracted from his set and compared with the remainder of his emails and all of the emails of the other 175 authors in the corpus, giving a total of 176 candidate authors. Similarity between the sample sets and the comparison sets is measured by Jaccard's similarity coefficient and in terms of all of the unigrams, bigrams, trigrams, four-grams, five-grams and six-grams that are found in both the sample and comparison texts, as a proportion of all n-grams of that size in the two sets combined. The expectation is that because Derrick authored the samples, the remainder of his emails should be most similar to them, and should therefore obtain the highest Jaccard similarity score of all 176 comparison sets and candidate authors in every test. In the analyses that follow, the accuracy, success and reliability of the different n-grams in attributing the various sample sets to Derrick are evaluated in two ways:

i. 'Raw attribution accuracy'. The most straightforward way of assessing success is by the number of correct attributions each n-gram achieves when applied to the different sample sizes. In any given test, if Derrick's remaining emails achieve the highest Jaccard score as compared to the samples of the other 175 authors, then attribution has been successful; if they do not, attribution has been unsuccessful.
ii. '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 Derrick

may not always achieve the highest Jaccard score in a given individual test, he may achieve Jaccard scores consistently high enough in each sample size so that he has the highest mean Jaccard score over the ten tests. In such a case, attribution is considered successful.

Results

Figure 4 compares the raw attribution accuracy of the different n-grams in identifying Derrick as the author of each of the ten samples for each size. First, with the smallest samples of 2% (9 emails, 55–145 tokens), performance is generally poor, with the best results being obtained by four-grams and five-grams which each have a success rate of 30%, attributing three of the ten samples to Derrick. Trigrams and six-grams only successfully attribute one sample each, while the shortest n-grams of unigrams and bigrams have no success at all. Whereas unigrams and bigrams continue to misattribute all ten 5% samples (23 emails, 156–398 tokens), n-grams from three to six words in length achieve a 60% success rate. When the samples reach 10% of Derrick's emails in size (46 emails, 350–622 tokens) trigrams and four-grams achieve 90% success rate, outperforming bigrams and five-grams which both achieve 80% success rates and six-grams and unigrams lagging behind at 40% and 30% success rates respectively. With the 15% sample sizes (70 emails, 537–875 tokens) bigrams, trigrams and four-grams perform perfectly, attributing all ten samples accurately to Derrick, while unigrams and six-grams also perform well with 90% accuracy and five-grams with 80%. Finally, by the time the samples to be attributed comprise 20% of Derrick's emails (93 emails, 763–1018 tokens), unigrams through to five-grams all achieve 100% accuracy, and six-grams follow closely behind with identifying Derrick as the author of nine of the ten samples.

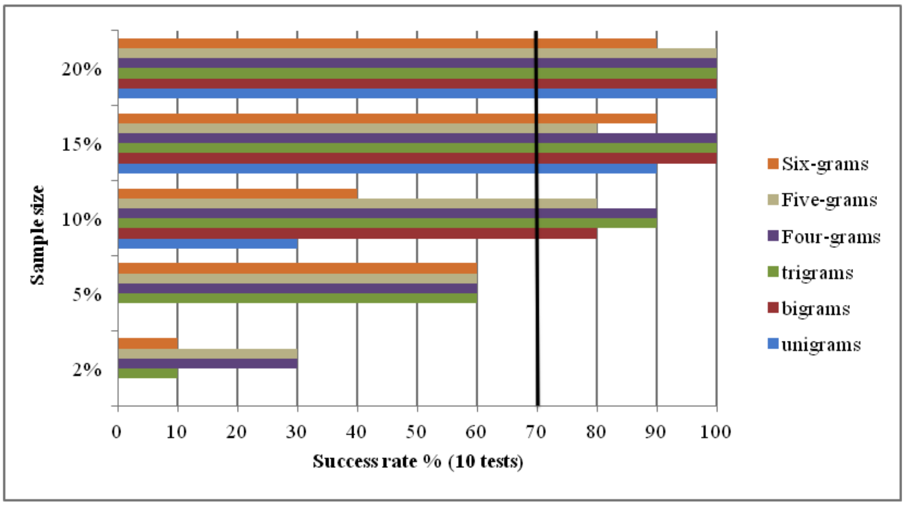


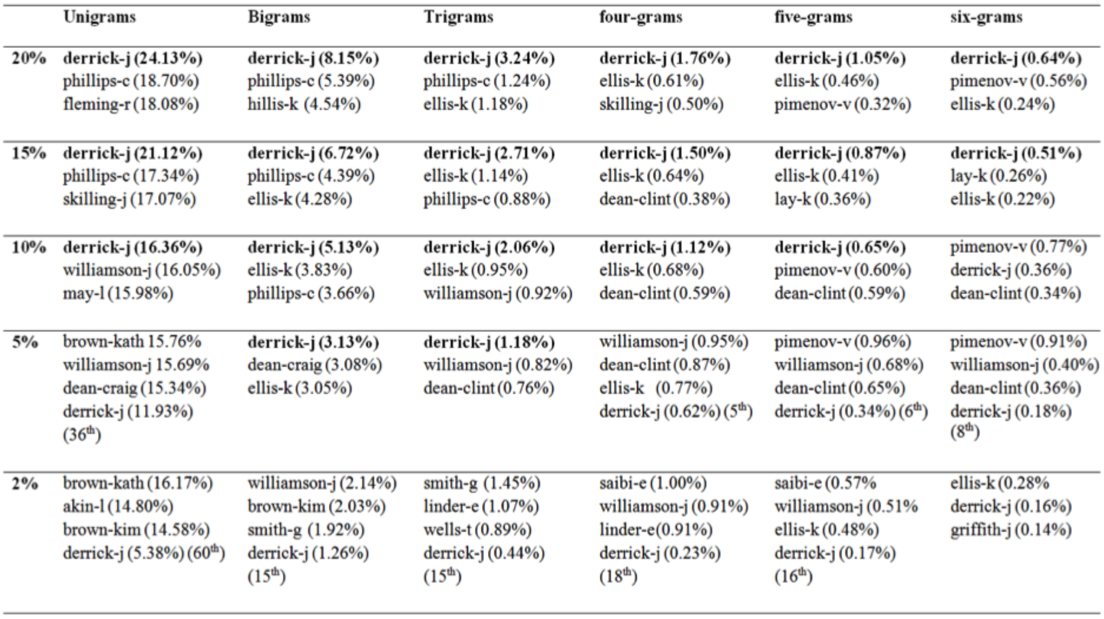
Figure 4. Accuracy rate of n-grams in successfully attributing samples of Derrick's emails.

Overall, these results are encouraging. Although there is not yet a consensus as to 'how good is good enough' when it comes to success and error rates in authorship research, 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'. If the less conservative 70% end of this threshold is applied, then everything to the right of the bold black line in Figure 4 can be considered 'successful'. This includes the results for bigrams, trigrams, four-grams and five-grams when applied to 10% samples, and the results for all of the six n-gram measures when tested on the 15% and 20% samples. What is most encouraging about the results in Figure 4 is that the number of accurate attributions beyond the 70% mark is not inconsiderable for the larger sample sizes. Despite not reaching the 70% threshold for the 2% and 5% samples, accuracy rates of 30% and 60% can be considered successful to some extent, particularly when the small sample sizes are taken into consideration (55–398 tokens). In fact, the three 2% samples successfully attributed by trigrams and four-grams are only 77, 84 and 109 tokens in size (see Table 1), which is remarkably small for computational authorship techniques. In a forensic case involving email, the analyst may have a very small number of disputed emails which they are asked to attribute, a few emails or maybe even only one. These small samples of between 77 and 109 tokens that have been successfully attributed to Derrick represent around nine of his emails. However, as noted above, he has a much lower words-per-email mean (12.9 tokens) than the rest of the authors in the corpus (41.74 tokens). Samples of these sizes would therefore comprise around two to three emails for the average Enron employee, which goes some way towards demonstrating the potential effectiveness of this method in forensic applications, especially as there are likely to be fewer candidate authors in any given case.

Based on the evidence provided by these results, the most accurate and reliable n-gram measure in identifying Derrick as the author of the samples is four-grams. Each n-gram measure underwent 50 tests, one for each sample (five sample sizes, ten samples each). Out of their 50 tests, four-grams correctly attributed 38 samples to Derrick (76%), closely followed by trigrams which successfully attributed 36 of the 50 samples (72%) and five-grams which attributed 35 (70%). On the other hand, unigrams and bigrams are the worst performers, successfully identifying Derrick as the author of only 22 (44%) and 28 (56%) samples respectively. While longer n-grams generally perform better than shorter ones, accuracy does not continue to rise with the length of the n-gram, with four-grams outperforming five- and six-grams. Although longer word strings are more likely to be distinctive of individual authors, it may be that these strings are less likely to be repeated by authors, while four-grams and even trigrams are more pervasive and consistent in a person's authorial style.

What these results do not take into account, though, is how closely Derrick comes to being identified as the author of the samples when he does not achieve the highest Jaccard score. Although he may not rank as being most similar to the sample in any one test, he may be the second or third most similar, or he may be in one-hundred-and-third place. Considering mean Jaccard scores of all 176 candidate authors across all ten tests of each sample size helps to circumvent this issue (Table 6). The table presents the authors with the three highest mean Jaccard scores for all ten tests for every sample size, for all six n-gram measures. With the 2% samples, Derrick does not achieve the highest mean Jaccard score over the ten samples using any of the six measures. He does, however, rank second using six-grams, and while unigrams perform very badly, ranking him at 60th of all 176 candidates, bigrams through to five-grams rank him between 15th and 18th. In a pool of 176 possible authors, n-grams between two and six words in length consistently rank Derrick as being in the top 10% (18/176) of authors. Given the very small size of these 2% samples, this is a promising result, and indicative of a better performance of the method than the raw attribution results (Figure 4) would suggest. With the 5%

samples, Derrick is successfully identified as the author of the ten samples using bigrams and trigrams, while four-, five- and six-grams rank him as 5th, 6th and 8th most similar respectively, or in other words within the top 5% of authors (8/176). While bigrams did not attribute a single 5% sample to Derrick (Figure 4) across the ten tests, the Jaccard scores obtained using bigrams were consistently high enough for him to score the highest mean of all the candidate authors. When the 10% samples were tested, Derrick achieved the highest mean Jaccard score using unigrams through to five-grams, with six-grams ranking him second. However, by the time the samples are 15% and 20% all six n-gram measures successfully identify Derrick as being most similar to the samples. Overall, with these mean Jaccard score results, if the performance of unigrams is disregarded, Derrick is never ranked outside the top 10% of the 176 candidate authors, and amongst these results bigrams and trigrams outperform the others as they are the only measures to successfully attribute the 5% samples to Derrick.



	Unigrams	**Bigrams**	**Trigrams**	**four-grams**	**five-grams**	**six-grams**
20%	**derrick-j (24.13%)** phillips-c (18.70%) fleming-r (18.08%)	**derrick-j (8.15%)** phillips-c (5.39%) hillis-k (4.54%)	**derrick-j (3.24%)** phillips-c (1.24%) ellis-k (1.18%)	**derrick-j (1.76%)** ellis-k (0.61%) skilling-j (0.50%)	**derrick-j (1.05%)** ellis-k (0.46%) pimenov-v (0.32%)	**derrick-j (0.64%)** pimenov-v (0.56%) ellis-k (0.24%)
15%	**derrick-j (21.12%)** phillips-c (17.34%) skilling-j (17.07%)	**derrick-j (6.72%)** phillips-c (4.39%) ellis-k (4.28%)	**derrick-j (2.71%)** ellis-k (1.14%) phillips-c (0.88%)	**derrick-j (1.50%)** ellis-k (0.64%) dean-clint (0.38%)	**derrick-j (0.87%)** ellis-k (0.41%) lay-k (0.36%)	**derrick-j (0.51%)** lay-k (0.26%) ellis-k (0.22%)
10%	**derrick-j (16.36%)** williamson-j (16.05%) may-l (15.98%)	**derrick-j (5.13%)** ellis-k (3.83%) phillips-c (3.66%)	**derrick-j (2.06%)** ellis-k (0.95%) williamson-j (0.92%)	**derrick-j (1.12%)** ellis-k (0.68%) dean-clint (0.59%)	**derrick-j (0.65%)** pimenov-v (0.60%) dean-clint (0.59%)	pimenov-v (0.77%) derrick-j (0.36%) dean-clint (0.34%)
5%	brown-kath 15.76% williamson-j 15.69% dean-craig (15.34%) derrick-j (11.93%) (36th)	**derrick-j (3.13%)** dean-craig (3.08%) ellis-k (3.05%)	**derrick-j (1.18%)** williamson-j (0.82%) dean-clint (0.76%)	williamson-j (0.95%) dean-clint (0.87%) ellis-k (0.77%) derrick-j (0.62%) (5th)	pimenov-v (0.96%) williamson-j (0.68%) dean-clint (0.65%) derrick-j (0.34%) (6th)	pimenov-v (0.91%) williamson-j (0.40%) dean-clint (0.36%) derrick-j (0.18%) (8th)
2%	brown-kath (16.17%) akin-l (14.80%) brown-kim (14.58%) derrick-j (5.38%) (60th)	williamson-j (2.14%) brown-kim (2.03%) smith-g (1.92%) derrick-j (1.26%) (15th)	smith-g (1.45%) linder-e (1.07%) wells-t (0.89%) derrick-j (0.44%) (15th)	saibi-e (1.00%) williamson-j (0.91%) linder-e(0.91%) derrick-j (0.23%) (18th)	saibi-e (0.57% williamson-j (0.51% ellis-k (0.48%) derrick-j (0.17%) (16th)	ellis-k (0.28% derrick-j (0.16%) griffith-j (0.14%)

Table 6. Authors with the highest mean Jaccard scores across all ten samples of each size.

Using this approach, we can also identify which of the other 175 Enron authors appear to be most similar to Derrick. Kaye Ellis appears in the top three authors 18 times in Table 6. Similarly, Joannie Williamson (10), Dean Clint (8), Cathy Phillips (7) and Vladi Pimenov (6) all repeatedly appear in the top three ranked authors, with samples being misattributed to Williamson and Pimenov on five occasions. The repeated appearance of these authors in the results attests to the consistency of the method; rather than the samples being misattributed to different authors each time, or dozens of different authors being scored as similar to Derrick, this way the method consistently identifies these authors as being the closest to him in stylistic terms. Combined with Derrick's consistent ranking in the top 10% of authors, these results provide evidence to suggest that besides being an effective method of attributing authorship of samples of varying sizes, the word n-gram approach may also be used to dramatically reduce the number of candidate authors in cases where the number of suspect writers is very large.

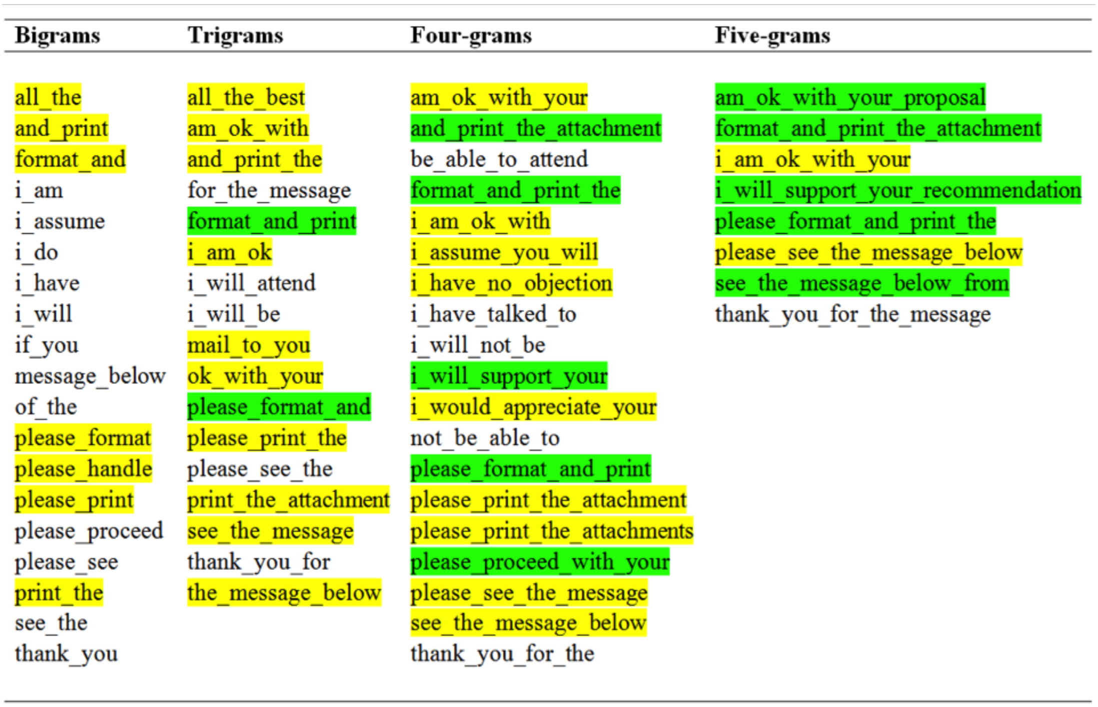
The first part of this case study (*Identifying Derrick's professional authorial style*) used a corpus-based stylistic approach to demonstrate how distinctive Derrick's collocation patterns are. Following from this, the results of this attribution experiment complement and confirm the findings of the stylistic analysis, in that they have shown word n-grams between one and six-words in length to be accurate and reliable in identifying Derrick as the author of the samples. In order to fully bridge the gap between the stylistic and computational approaches, the third and final part of the case study examines the actual n-grams that are responsible for the accurate attribution of authorship in these experiments.

Derrick's discriminating n-grams

One of the main advantages of *Jangle* is that, as well as running comparisons and calculating Jaccard scores, the program also displays the actual n-grams operating behind the statistics and accounting for the Jaccard scores. This kind of information about which specific linguistic choices were most important in attribution tasks is often not made available in stylometric studies. In this particular attribution case, this facility allows us to pinpoint the n-grams that were most useful in identifying Derrick as the author of the disputed samples, and to observe any recurrent patterns (textbites), and compare these with those identified in the corpus stylistic analysis. The examination here focuses on the bigrams, trigrams, four-grams and five-grams which contributed to the accurate attribution of the 5% sample sets. Although bigrams were not successful in attributing the individual 5% samples, they were successful in scoring Derrick the highest mean Jaccard score across all ten tests with this sample size, and for this reason the shared bigrams in all ten 5% samples are examined here. In contrast, the trigrams, four-grams and five-grams were successful in attributing six of the ten individual 5% samples, and the n-grams considered here are those shared between the samples and the remainder of Derrick's emails in these successful tests.

In total there were 311 different bigrams that were found in both the sample set and the remainder of Derrick's emails across the ten 5% sample tests, and there were 122 trigrams, 64 four-grams and 28 five-grams that were found in both sets of emails in the six or seven successful tests. One major criticism of using content words or combinations of content words in authorship analysis is that they are indicative of topic rather than of authorship. Indeed, there is a risk that Derrick's samples have been attributed to him on the basis that the emails in the two different sets made reference to the same topics, such as people, place or company names. However, only 57 (18%) of the 311 bigrams could be considered to refer to topic specific entities, for example *jim derrick*, *steph please*, *enron letterhead*, *southwestern legal*, and *the litigation*. Similarly, 24 of the 122 trigrams (19.7%), 19 of the 64 four-grams (29.7%) and 11 of the 28 five-grams (39.3%) can be considered as being topic-dependent in this way. Such n-grams have been removed from the analysis that follows. Thus, although the proportion of topic-dependent n-grams gradually increases as the length of the n-gram is extended, the majority (61%-82%) of the useful n-grams are topic independent; that is, they are not shared between the samples and the remainder of Derrick's emails simply because Derrick is talking about the same things or about/to the same people.

Table 7 presents bigrams which appeared in both the sample sets and Derrick's remaining emails in five or more tests, and the tri-, four-, and five-grams which were shared between Derrick's sample and his remaining emails in three or more successful tests. The n-grams displayed in the table collectively represent a pool of the most characteristic and discriminatory n-gram textbites repeatedly used by Derrick. They have all contributed to the successful attribution of his 5% sample sets which, varying between 156 and 398 words in length, are the smallest sample sets accurately attributed in the experiment besides three 2% samples (*Results* section).



Bigrams	Trigrams	Four-grams	Five-grams
all_the	all_the_best	am_ok_with_your	am_ok_with_your_proposal
and_print	am_ok_with	and_print_the_attachment	format_and_print_the_attachment
format_and	and_print_the	be_able_to_attend	i_am_ok_with_your
i_am	for_the_message	format_and_print_the	i_will_support_your_recommendation
i_assume	format_and_print	i_am_ok_with	please_format_and_print_the
i_do	i_am_ok	i_assume_you_will	please_see_the_message_below
i_have	i_will_attend	i_have_no_objection	see_the_message_below_from
i_will	i_will_be	i_have_talked_to	thank_you_for_the_message
if_you	mail_to_you	i_will_not_be	
message_below	ok_with_your	i_will_support_your	
of_the	please_format_and	i_would_appreciate_your	
please_format	please_print_the	not_be_able_to	
please_handle	please_see_the	please_format_and_print	
please_print	print_the_attachment	please_print_the_attachment	
please_proceed	see_the_message	please_print_the_attachments	
please_see	thank_you_for	please_proceed_with_your	
print_the	the_message_below	please_see_the_message	
see_the		see_the_message_below	
thank_you		thank_you_for_the	

Table 7. Useful n-grams in the successful attribution of Derrick's 5% email samples (green highlights unique to Derrick; yellow indicate Derrick is top frequency author among shared n-grams).

One recurring n-gram pattern that emerges from these results is strings which begin with *I*, for example the trigram *I have talked*, the four-grams *I assume you will*, *I would appreciate your*, and the five-grams *I am ok with your*, *I have no objection to*, and *I will support your recommendation* (Note the appearance of *I would appreciate your* in Example 7 and in Figure 2 in the stylistic analysis). All of these n-grams are productive in discriminating Derrick's email style from that of the other candidate authors, and useful in correctly attributing his samples.

However, the most obvious pattern in these results is the recurrence of *please* and *thank you* initial n-grams. These appear from bigrams such as *please print*, *please see* and *thank you* to four-grams such as *please print the attachment* and five-grams such as *please see the message below* and *thank you for the message*. What the stylistic analysis identifies that the computational analysis does not, is the co-selection of *please* with *Thank you*. This is a limitation of the search for n-grams within punctuated sentences. However, the corpus analysis complements this and the within-sentence rule ensures that meaningful strings are preserved. Other distinctive patterns are the word n-grams which are contained within the longer politeness strings, for example bigrams such as *format and* and *the message* to four-grams and five-grams such as *see the message below* and *format and print the attachment*. These results complement and confirm what was found in the corpus stylistic analysis in the section *Identifying Derrick's professional authorial style*. The qualitative stylistic analysis in the first instance found that Derrick was remarkably formulaic in his use of politeness structures, and that there are a number of n-grams that are distinctive of him in the Enron corpus. In turn, this section has found that it is exactly these n-grams that are most influential in the statistical attribution experiment.

Using these results, we can return again to a corpus approach to identify how common or rare these textbites are in the rest of the Enron corpus. In other words, are these n-grams found in the emails of the other 175 Enron authors? Those highlighted in green in Table 7 are those that are unique to Derrick; when searched for using *Wordsmith Tools*, the only results returned are from his emails. Taking results from across the four lengths of n-gram we can identify five n-gram textbites that are unique to him: *please format and print the attachment*, *I will support your recommendation*, *please proceed with your*, *am OK with your proposal* and *see the message below from*. In addition, those in yellow are those that are shared with other authors in the corpus, but Derrick is the top frequency user per 1,000 words of text. For example, *please print the* is used by twelve authors in the corpus, but Derrick uses it with a greater relative frequently than the other eleven. With the green and yellow n-grams combined, we have a pool of textbites that are distinctive of Derrick and characterise his authorial style when compared with the other authors in the Enron population. These findings show the value of returning to the corpus approach in validating the n-grams further, and provide an explanation for why these particular n-grams were useful in the successful attribution of Derrick's samples.

The methodological implications of this successful triangulation of approaches are considerable. On the one hand, the statistical attribution task confirms and supports the reliability of the stylistic results in terms of the distinctiveness and usefulness of Derrick's *please*-initial n-grams. At the same time, the stylistic analyses in the first half of the case study offers a clear linguistic explanation for the statistical and computational results in terms of why n-grams are an accurate and useful linguistic feature for distinguishing an individual's writing style and attributing authorship. Throughout his life, Derrick has amassed a unique series of linguistic and professional experiences culminating in being a legal educator and practitioner, and word collocations and patterns are stored and primed in his mind based on these unique experiences. As such, when Derrick finds himself within the relevant communicative context with regard to purpose and recipient of the email, he accesses and produces stored and frequently used collocation patterns which, in turn, prove to be distinctive of his email style. Moreover, he produces such patterns regularly and repeatedly, reinforcing the priming and making them increasingly habitual. The result is that a word n-gram approach can be used not only to identify and isolate a number of n-gram textbites that distinguish his professional email style from that of other employees in the same company, but also to successfully identify him as the author of text samples, including some as small as 77, 84 and 109 tokens.

Conclusion

Turell (2010) notes the important role that research studies in authorship attribution play in benefiting the performance of expert witnesses in cases of disputed authorship. In the 21st century forensic linguistic casework increasingly involves texts created in online modes, including email. This research, therefore, provides much-needed empirical results that might provide knowledge for expert witness reports in the future. We now know that focusing on one style-marker, *please*, and one speech act, politely encoded directives, can produce both stylistically and statistically revealing results, even though we know that *please* is one of the most frequent lexical items in email and that directives are one of the most important speech acts in emails. Even so, it is possible to find distinctive uses of these items and habitual behaviours within a large pool of authors. In most authorship casework reports focus on a whole range of different linguistic style markers (lexical, syntactic, orthographical, etc.), but this research has shown that this is not necessary, when stylistic analysis is combined with statistical Jaccard similarity testing. This reinforces the benefits of a triangulation of approaches: corpus stylistic, statistical, computational, and case study, following the call by Solan and Tiersma (2004: 463) to test and validate methods "with proven reliability in determining authorship" in order that the legal system finds the expert evidence of forensic linguists acceptable.

The statistical attribution method illustrated and tested in the *Attribution task* section has an accuracy rate of 100% for larger samples, and promising results for very small samples. It allows us to reduce a mass of data to meaningful "textbites" that can characterise and identify authors, in this study one author: James Derrick. The combination of corpus stylistic, computational, and statistical methods, then coming back to the corpus to verify results, produces a set of unique and highly distinctive textbites for Derrick. This also shows us what a corpus like the Enron corpus can enable researchers to do in the forensic linguistic field. Having a dataset of authors as large as 176, allows us to explore author email style, understanding more fully how writers who are (socio)linguistically close in many ways, writing within the same text type, in the same company, at the same time, can do so in ways that are distinctive. Even though they are selecting the same word (*please*), the collocations and co-selections they make vary to form recurrently different n-grams that are both stylistically and statistically observable and quantifiable. The identification of these n-gram textbites moves us closer to the elusive concept of idiolect.

Acknowledgments

The authors would like to thank David Woolls of CFL Software Limited (www.cflsoftware.com) for supplying us with the *Jangle* program, without which this research would not have been possible in its current form. Automising the process of producing random samples and calculating the Jaccard similarity at the touch of a button is something (forensic) linguists can only wonder at. Then giving the linguist the raw outputs of the program, in addition, so that we can verify and analyse the n-grams that have produced the results, puts the linguistic evidence at the forefront of the computational technique.

David Wright's contribution was supported by an AHRC studentship: Grant number: AH/J502686/1.

References

Ahrens, F. (2006). The first thing we do, let's hire all the lawyers. *The Washington Post*, http://blog.washingtonpost.com/enron/2006/04/the_first_thing_we_do_lets_hir_1.html.

Ancu, M. (2011). From soundbite to textbite: Election '08 comments on twitter. In J. Hendricks and L. L. Kaid, Eds., *Techno Politics in Presidential Campaigning. New Voices, New Technologies, and New Voters*, 11–21. London: Routledge.

Argamon, S. and Koppel, M. (2010). The rest of the story: Finding meaning in stylistic variation. In S. Argamon, K. Burns and S. Dubnov, Eds., *The Structure of Style: Algorithmic Approaches to Understanding Manner and Meaning*, 79–112. London: Springer.

Argamon, S. and Koppel, M. (2013). A systemic functional approach to automated authorship analysis. *Journal of Law and Policy*, 21(2), 299–316.

Argamon, S. and Levitan, S. (2005). Measuring the usefulness of function words for authorship attribution. In *Proceedings of ACH/ALLC Conference*, 1–3: University of Victoria, BC, Association for Computing and the Humanities.

Austin, J. L. (1962). *How to Do Things with Words.* Oxford: Clarendon Press.

Barlow, M. (2010). Individual usage: A corpus-based study of idiolects. Paper presented at *34th International LAUD Symposium.* Landau, Germany.

Barlow, M. (2013). Exemplar theory and patterns of production. Paper presented at *Corpus Linguistics 2013.* Lancaster, UK.

Baron, N. S. (1998). Letters by phone or speech by other means: The linguistics of email. *Language and Communication*, 18(2), 133–170.

Bax, P. (1986). How to assign work in an office. A comparison of spoken and written directives in American English. *Journal of Pragmatics*, 10, 673–692.

Bennell, C. and Jones, N. J. (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.

Bou-Franch, P. (2011). Openings and closings in Spanish email conversations. *Journal of Pragmatics*, 43(6), 1772–1785.

Bou-Franch, P. and Lorenzo-Dus, N. (2008). Natural versus elicited data in cross-cultural speech act realisation: The case of requests in Peninsular Spanish and British English. *Spanish in Context*, 5(2), 246–277.

Burrows, J. (2002). "Delta:" A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3), 267–287.

Burrows, J. (2005). Andrew Marvell and the "Painter Satires": A computational approach to their authorship. *Modern Language Review*, 100(2), 281–297.

Chaski, C. 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, C. E. (2007). Multilingual forensic author identification through n-gram analysis. Paper presented at the *8th Biennial Conference on Forensic Linguistics/Language and Law.* Seattle, WA.

Chejnová, P. (2014). Expressing politeness in the institutional e-mail communications of university students in the Czech Republic. *Journal of Pragmatics*, 60, 175–192.

Cohen, W. W. (2009). Enron email dataset. [online] http://www.cs.cmu.edu/ enron/ (Accessed November 2010).

Cotterill, J. (2010). How to use corpus linguistics in forensic linguistics. In A. O'Keefe and M. McCarthy, Eds., *The Routledge Handbook of Corpus Linguistics*, 578–590. London: Routledge.

Coulthard, M. (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, M. (2004). Author identification, idiolect, and linguistic uniqueness. *Applied Linguistics*, 24(4), 431–447.

Coulthard, M. (2013). On admissible linguistic evidence. *Journal of Law and Policy*, 21(2), 441–466.

Coulthard, M., Grant, T. and Kredens, K. (2011). Forensic linguistics. In R. Wodak, B. Johnstone and P. Kerswill, Eds., *The SAGE Handbook of Sociolinguistics*, 531–544. London: Sage.

Coyotl-Morales, R. M., Villaseñor-Pineda, L., Montes-y-Gómez, M. and Rosso, P. (2006). Authorship attribution using word sequences. In *Proceedings of the 11th Iberoamerican Congress on Pattern Recognition*, 844–853, Berlin: Springer.

Creamer, G., Rowe, R., Hershkop, S. and Stolfo, S. J. (2009). Automated social hierarchy detection through email network analysis. In H. Zhang, M. Spiliopoulou, B. Mobasher, C. L. Giles, A. McCallum, O. Nasraoui, J. Srivastava and J. Yen, Eds., *Proceedings of the 9th International Workshop on Mining Web Data, WEBKDD 2007, and the 1st International Workshop on Social Network Analysis, SNA-KDD 2007.*

Culwin, F. and Child, M. (2010). Optimising and automating the choice of search strings when investigating possible plagiarism. In *Proceedings of the 4th International Plagiarism Conference, Newcastle, UK [online]*, http://www.plagiarismadvice.org/research-papers/item/optimising-and-automating-the-choice-of-search-strings-when-investigating-possible-plagiarism (Accessed July 2013).

Davies, M. (2012). The Corpus of Contemporary American English: 450 million words, 1990-present [online]. http://corpus.byu.edu/coca/.

Deng, F., Siersdorfer, S. and Zerr, S. (2012). Efficient Jaccard-based Diversity Analysis of Large Document Collections. In *Proceedings of the 21st ACM international Conference on Information and Knowledge Management (CIKM '12)*, 1402–1411.

Duthler, K. W. (2006). The politeness of requests made via email and voicemail: Support for the hyperpersonal model. *Journal of Computer-Mediated Communication*, 11(2), 500–521.

Economidou-Kogetsidis, M. (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.

FERC, F. E. R. C. (2013). Information released in Enron investigation. [online] http://www.ferc.gov/industries/electric/indus-act/wec/enron/info-release.asp.

Firth, J. R. (1957). A synopsis of linguistic theory 1930-1955. In F. Palmer, Ed., *Selected Papers of J.R. Firth 1952-1959*, 168–205. London: Longman.

Flowerdew, L. (2004). The argument for using English specialized corpora to understand academic and professional language. In U. Connor and T. A. Upton, Eds., *Discourse in the Professions: Perspectives from Corpus Linguistics*, volume 16, 11–33. Amsterdam: John Benjamins.

Gains, J. (1999). Electronic mail – a new style of communication or just a new medium? An investigation into the text features of e-mail. *English for Specific Purposes*, 18(1), 81–101.

Gimenez, J. C. (2000). Business e-mail communication: Some emerging tendencies in register. *English for Specific Purposes*, 19(3), 237–251.

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

Grant, T. (2008). Approaching questions in forensic authorship analysis. In J. Gibbons and M. T. Turell, Eds., *Dimensions of Forensic Linguistics*, 215–229. Amsterdam: John Benjamins.

Grant, T. (2010). Txt 4n6: Idiolect free authorship analysis? In M. Coulthard and A. Johnson, Eds., *The Routledge Handbook of Forensic Linguistics*, 508–522. London: Routledge.

Grant, T. (2013). Txt 4n6: Method, consistency, and distinctiveness in the analysis of SMS text messages. *Journal of Law and Policy*, 21(2), 467–494.

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

Gries, S. T. (2013). 50-something years of work on collocations: What is or should be next.... *International Journal of Corpus Linguistics*, 18(1), 137–165.

Grieve, J. (2007). Quantitative authorship attribution: An evaluation of techniques. *Literary and Linguistic Computing*, 22(3), 251–270.

Hirst, G. and Feiguina, O. (2007). Bigrams of syntactic labels for authorship discrimination of short texts. *Literary and Linguistic Computing*, 22(4), 405–417.

Hoey, M. (2005). *Lexical Priming: A new theory of words and language.* London: Routledge.

Hoover, D. L. (2002). Frequent word sequences and statistical stylistics. *Literary and Linguistic Computing*, 17(2), 157–180.

Hoover, D. L. (2003). Frequent collocations and authorial style. *Literary and Linguistic Computing*, 18(3)(3), 261–228.

Hoover, D. L. (2004). Testing Burrows's Delta. *Literary and Linguistic Computing*, 19(4), 453–475.

Hoover, D. L. (2009). Word frequency, statistical stylistics and authorship attribution. In D. Archer, Ed., *What's in a word list? Investigating word frequency and keyword extraction*, 35–51. Surrey: Ashgate.

Hoover, D. L. (2010). Authorial style. In D. McIntyre and B. Busse, Eds., *Language and Style*, 250–271. Basingstoke: Palgrave MacMillan.

Izsak, C. and Price, A. R. (2001). Measuring beta-diversity using a taxonomic similarity index, and its relation to spatial scale. *Marine Ecology Progress Series*, 215, 69–77.

Jaccard, P. (1912). The distribution of the flora in the alpine zone. *The New Phytologist*, 11(2), 37–50.

Johnson, A. and Woolls, D. (2009). Who wrote this? The linguist as detective. In S. Hunston and D. Oakey, Eds., *Introducing Applied Linguistics: Concepts and Skills*, 111–118. London: Routledge.

Juola, P. (2008). *Authorship Attribution. Foundations and Trends in Information Retrieval.* Delft: NOW Publishing.

Juola, P. (2013). Stylometry and immigration: A case study. *Journal of Law and Policy*, 21(2), 287–298.

Knox, J. (2007). Visual-verbal communication on online newspaper home pages. *Visual Communication*, 6(1), 19–53.

Koppel, M., Schler, J. and Argamon, S. (2009). Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, 60(1), 9–26.

Koppel, M., Schler, J. and Argamon, S. (2011). Authorship attribution in the wild. *Language Resources and Evaluation*, 45(1), 83–94.

Koppel, M., Schler, J. and Argamon, S. (2013). Authorship attribution: What's easy and what's hard? *Journal of Law and Policy*, 21(2), 317–332.

Koppel, M., Schler, J., Argamon, S. and Messeri, E. (2006). Authorship attribution with thousands of candidate authors. In *Proceedings of the 29th ACM SIGIR Conference on Research and Development on Information Retrieval*, Seattle, Washington.

Kredens, K. (2002). Towards a corpus-based methodology of forensic authorship attribution: A comparative study of two idiolects. In B. Lewandowska-Tomaszczyk, Ed., *PALC'01: Practical Applications in Language Corpora*, 405–437. Frankfurt am Main: Peter Lang.

Kredens, K. and Coulthard, M. (2012). Corpus linguistics in authorship identification. In P. Tiersma and L. Solan, Eds., *The Oxford Handbook of Language and Law*, 504–516. Oxford: Oxford University Press.

Labbé, D. (2007). Experiments on authorship attribution by intertextual distance in English. *Journal of Quantitative Linguistics*, 14(1), 33–80.

Lampert, A., Dale, R. and Paris, C. (2008). The nature of requests and commitments in email messages. In *Proceedings of the AAAI Workshop on Enhanced Messaging 2008*, 42–47.

Lan, L. (2000). Email: A challenge to Standard English? *English Today*, 16(4), 23–29.

Larner, S. (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.

Love, H. (2002). *Attributing Authorship: An Introduction.* Cambridge: Cambridge University Press.

Luchjenbroers, J. and Aldridge-Waddon, M. (2011). Paedophiles and politeness in email communications: Community of practice needs that define face-threat. *Journal of Politeness Research: Language, Behaviour, Culture*, 7(1), 21–42.

Luyckx, K. and Daelemans, W. (2011). The effect of author set size and data size in authorship attribution. *Literary and Linguistic Computing*, 26(1), 35–55.

Manasse, M. (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.

Markson, L., Woodhams, J. and Bond, J. W. (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.

Mazzoleni, G. and Schulz, W. (1999). "Mediatization" of politics: A challenge for democracy? *Political Communication*, 16(3), 247–261.

Merrison, A. J., Wilson, J. J., Davies, B. L. and Haugh, M. (2012). Getting stuff done: Comparing e-mail requests from students in higher education in Britain and Australia. *Journal of Pragmatics*, 44(9), 1077–1098.

Mollin, S. (2009). "I entirely understand" is a Blairism: The methodology of identifying idiolectal collocations. *International Journal of Corpus Linguistics*, 14(3), 367–392.

Mosteller, F. and Wallace, D. L. (1964). *Inference and Disputed Authorship: The Federalist.* Reading, MA: Addison-Wesley Publishing Company Inc.

Najeeb, Z. M., Maros, M. and Nor, N. F. M. (2012). Politeness in e-mails of Arab students in Malaysia. *GEMA Online Journal of Language Studies*, 12(1), 125–145.

Nattinger, J. R. and DeCarrico, J. S. (1992). *Lexical Phrases and Language Teaching.* Oxford: Oxford University Press.

Naumann, F. and Herschel, M. (2010). An introduction to duplicate detection. *Synthesis Lectures on Data Management*, 2(1), 1–87.

Nerlich, B., Forsyth, R. and Clarke, D. (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.

Nini, A. and Grant, T. (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.

Pace-Sigge, M. (2013). *Lexical Priming in Spoken English Usage.* New York: Palgrave MacMillan.

Pottier, J., Dubuis, A., Pellissier, L., Maiorano, L., Rossier, L., Randin, C. F., Vittoz, P. and Guisan, A. (2013). The accuracy of plant assemblage prediction from species distribution models varies along environmental gradients. *Global Ecology and Biogeography*, 22(1), 52–63.

Priebe, C. E., Conroy, J. M., Marchette, D. J. and Park, Y. (2005). Scan Statistics on Enron Graphs. *Computational and Mathematical Organization Theory*, 11(3), 229–247.

Queralt, S. and Turell, M. T. (2012). Testing the discriminatory potential of sequences of linguistic categories (n-grams) in Spanish, Catalan and English corpora. In *Paper presented at the Regional Conference of the International Association of Forensic Linguists 2012*, Kuala Lumpur (Malaysia): University of Malaya.

Rajaraman, A. and Ullman, J. D. (2011). *Mining of Massive Datasets.* Cambridge: Cambridge University Press.

Rico-Sulayes, A. (2011). Statistical authorship attribution of Mexican drug trafficking online forum posts. *The International Journal of Speech, Language and the Law*, 18(1), 53–74.

Sanderson, C. and Guenter, S. (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*, 482–491, Morristown, NJ: Association for Computational Linguistics.

Savoy, J. (2012). Authorship attribution: A comparative study of three text corpora and three languages. *Journal of Quantitative Linguistics*, 19(2), 132–161.

Schmitt, N., Grandage, S. and Adolphs, S. (2004). Are corpus-derived recurrent clusters psycholinguistically valid? In N. Schmitt, Ed., *Formulaic Sequences: Acquisition, Processing and Use*, 127–151. Amsterdam: John Benjamins Publishing Company.

Scott, M. (2008). *Wordsmith Tools version 5.* Liverpool: Lexical Analysis Software.

Scott, M. (2010). *Wordsmith Tools Help.* Liverpool: Lexical Analysis Software.

Searle, J. R. (1969). *Speech Acts: An Essay in the Philosophy of Language.* London: Cambridge University Press.

Sherblom, J. (1988). Direction, function and signature in electronic mail. *The Journal of Business Communication*, 25(4), 39–54.

Sinclair, J. M. (1991). *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

Solan, L. (2013). Intuition versus algorithm: The case for forensic authorship attribution. *Journal of Law and Policy*, 21(2), 551–576.

Solan, L. and Tiersma, P. M. (2004). Author identification in American Courts. *Applied Linguistics*, 25(4), 448–465.

Stamatatos, E. (2008). Author identification: Using text sampling to handle the class imbalance problem. *Information Processing and Management*, 44(2), 790–799.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3), 538–556.

Stamatatos, E. (2013). On the robustness of authorship attribution based on character n-gram features. *Journal of Law and Policy*, 21(2), 421–440.

Stubbs, M. (2001). *Words and Phrases: Corpus Studies of Lexical Semantics.* Oxford: Blackwell.

Tang, Z., Fang, J., Chi, X., Feng, J., Liu, Y., Shen, Z., Wang, X., Wang, Z., Wu, X., Zheng, C. and Gaston, K. J. (2013). Patterns of plant beta-diversity along elevational and latitudinal gradients in mountain forests of China. *Ecography*, 35(12), 1083–1091.

Turell, M. T. (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.

Van Den Eynden, N. (2012). Politeness and gender in Belgian organisational emails. In P. Gillaerts, E. de Groot, S. Dieltjens, P. Heynderickx and G. Jacobs, Eds., *Researching Discourse in Business Genres: Cases and corpora*, 33–52. Bern: Peter Lang.

van Halteren, H., Baayen, H., Tweedie, F., Haverkort, M. and Neijt, A. (2005). New machine learning methods demonstrate the existence of a human stylome. *Journal of Quantitative Linguistics*, 12(1), 65–77.

Waldvogel, J. (2007). Greetings and closings in workplace email. *Journal of Computer-Mediated Communication*, 12(2), 456–477.

Woodhams, J., Hollin, C. and Bull, R. (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.

Woolls, D. (2012). Description of CFL extraction routines for CMU Enron Sent email database. [Online] http://www.cflsoftware.com/CFL_CMU_Enron_Sent_email_Extraction.mht.

Woolls, D. (2013). *CFL Jaccard n-gram Lexical Evaluator. Jangle* version 2. CFL Software Limited.

Wray, A. (2002). *Formulaic Language and the Lexicon.* Cambridge: Cambridge University Press.

Wray, A. (2008). *Formulaic Language: Pushing the boundaries.* Oxford: Oxford University Press.

Wright, D. (2012). Existing and innovative techniques in authorship analysis: Evaluating and experimenting with computational approaches to ‘big data’ in the Enron Email Corpus. In *Paper presented at The 3^rd^ European Conference of the International Association of Forensic Linguists*, Porto, Portugal.

Wright, D. (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.

Yin, R. K. (2009). *Case Study Research Design and Methods.* London: Sage Ltd, 4th ed.

Zheng, R., Li, J., Chen, H. and Huang, Z. (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.