This is a post-review, pre-publication version of the paper: 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.

Stylistic variation within genre conventions in the Enron email corpus: Developing a text-sensitive methodology for authorship research

1 Introduction

A theory of idiolect is central to authorship analysis (Coulthard 2004:431) and over recent years there has been discussion of how idiolect can be best conceptualised for use in forensic authorship analysis (Kredens 2002; Grant 2010; Turell 2010). This paper takes a genre-focused approach to empirically investigating idiolect and linguistic individuality. Using email data from the former American energy company Enron, analysis focuses on author-distinctive variation within two conventions of the email genre – greetings and farewells. First, a number of author-distinctive greeting and farewell forms are found to distinguish between small set of four Enron traders. Then, using likelihood ratios, the rarity, expectancy and distinctiveness of the features identified are tested against a 126-author 'Enron Sent Email Author Reference Corpus'. Likelihood ratios are used commonly in medical research and practice (e.g. McGee 2002) and Coulthard (2010) shows how they can be applied to forensic authorship analysis.

2 Idiolect and authorship analysis

Coulthard (2004:431) states that 'the linguist approaches the problem of questioned authorship from the theoretical position that every native speaker has their own distinct and individual version of the language they speak and write, their own *idiolect*'. Kredens (2002:405), while acknowledging that there is 'universal agreement that an individual's linguistic repertoire is in some way distinct', also highlights that 'this claim has not thus far been supported by empirical research'. In his paper, Kredens compares the idiolects of two English musicians – Robert Smith and Patrick Morrissey – through a quantitative analysis of a range of linguistic features, with results proving 'some degree of idiolectal variation'. There are a small number of empirical studies of this kind into idiolectal variation. Yet, although forensic linguistics – and authorship analysis in particular – is the field of linguistics which arguably relies most heavily on the theory of idiolect, the empirical research into its existence and accessibility is emerging from other fields, such as corpus linguistics (e.g. Mollin 2009; Barlow 2010) and sociolinguistics (e.g. Kuhl 2003; Johnstone 2009), with such studies showing that that it is possible to make some substantiated claims about linguistic individuality and uniqueness.

The reasons behind the dearth of idiolect research from the forensic standpoint are clear. As Coulthard (2004:432) states, the collection and analysis of sufficient data to make any strong claims regarding a 'linguistic fingerprint' is 'impractical if not impossible'. Similarly, Turell (2010:217) argues that for idiolectal evidence to be useful in a forensic context requires achieving 'a *Base Rate Knowledge* of all possible realizations of an individual's idiolect, in spite of being a daunting endeavour of measuring the impossible'. The result of this difficulty is such that a theory of idiolect may remain too abstract or idealised to be of practical use to the forensic linguist. Consequently, the current situation is one in which a theoretical conversation is beginning with regards to the status of idiolect in authorship analysis. Grant (2010:522) suggests that a theory of idiolect may not necessarily be required for successful attribution of authorship, arguing that 'consistency and pair-wise

distinctiveness [between authors] are matters of empirical observation upon which forensic authorship analysis can rely', and that such authorship analysis 'can be considered idiolect free, or at least idiolect light'. Similarly, Turell (2010:216–8) offers insight into the theoretical controversies surrounding the theory, and, in the same way as Kredens (2002:405), suggests that perhaps 'idiolects have been accepted as *a priori* constructs and their existence assumed too readily'. Moreover, Turell (2010:217) introduces the notion of 'idiolectal style' as a less idealised construct, and one which can 'be more relevant to forensic authorship contexts', in that it is not primarily concerned with the language system an individual has, but how this system is *used* by a particular speaker/writer in ways which appear to be individual and unique. Although not from forensic linguistics, De Beaugrande (1998:28) in his discussion of the 'real and ideal' in linguistics, discusses *discoursolect*, or 'the current and episodic online system', as a less idealised 'specification' of idiolect. Both idiolectal style and discoursolect, as more accessible conceptualisations of idiolect, relate to language *in use* at a particular time. Therefore, given the relationship between genre and language use (Hymes 1974; Halliday and Hasan 1989; Biber 1993; Biber and Conrad 2009), the linguistic choices comprising a person's discoursolect or idiolectal style displayed in any one text, will depend on the context of writing and the genre in which they are writing.

In authorship casework and research, forensic linguists have productively made use of the ways in which individuals' linguistic choices vary distinctively within the conventions of the given genre. In his forensic analysis of text messages in the Jenny Nicholl murder case, Coulthard highlighted nine linguistic features which discriminated between the two authors in question: the victim and the murder suspect, her lover David Hodgson (See Grant 2010). A number of these features were abbreviations, (lack of) space between words, spelling – such as the use of *cu*, *cya*, *fone*, and no spaces between words either side of *2*, as in *booked2go bowling* – all of which are linguistic features typical of text messages (Crystal 2008). MacLeod and Grant (2012:218) have developed a sophisticated multi-level feature categorisation system for Tweets based on features characteristic of computer mediated communication and short form messages (Crystal 2001; Smith, Spencer and Grant 2009). The situation is much the same for email; several authorship studies have found that linguistic variation within generic email conventions such as greetings and sign-offs have been useful in determining authorship (de Vel, Anderson, Corney and Mohay 2001; Abbasi and Chen 2005; Turell 2010:225–6). Therefore, analysing the linguistic behaviour of an individual within the conventions of specific genres may be a useful point of entry for investigation into their idiolect. In fact, one may go as far to say that an individual's idiolect is a combination of numerous *genre-lects*, the former of which is much more difficult to analyse than the latter. To that end, this paper empirically investigates individuals' idiolects through measuring how distinctive an author's stylistic choices can be within even as few as two elements of the email genre – greetings and farewells.

3 Data: a corpus-based approach to authorship analysis

The data in this study are emails written by employees of the former American energy commodities and services company Enron. As part of the Federal Energy Regulatory Commission's (FERC) legal investigation into the company's illegal accounting practices, the email data of around 150 Enron employees, or 'custodians', containing approximately half a million emails, was made publically available online. The source of the data for this study is that provided by Carnegie Mellon University (CMU) (Cohen 2009). The data was extracted, prepared and designed for the purposes of this research by Woolls (2012) using specialised data extraction software. The extraction process created two sub-corpora from the

CMU dataset: 'The Trader Sent Corpus', which contains the emails of four traders, and the 'Enron Sent Email Author Reference Corpus', which contains the emails of a further 126 authors.

3.1 The 'Trader Sent Corpus'

The first stage of analysis in this study examines the distinctiveness of linguistic choices within a subset of four authors who are all traders: John Arnold, Chris Germany, John Lavorato and Andy Zipper, and who, combined, form a corpus comprising 2,622 emails and 86,902 words (Table 1).

Table 1: Composition of the Trader Sent Corpus for this study

	Arnold	**Germany**	**Lavorato**	**Zipper**	**Total**
Emails	632	1,339	405	246	2,622
Words	20,890	47,543	9,721	8,748	86,902
Average email length (words)	33	36	24	36	33

For the four authors in this 'Trader Sent Corpus', all of the emails from their 'sent', 'sent_items', or '_sent_items' folder were extracted and have been included, depending on which folder contained the most emails. Emails which were written by someone else on behalf of the traders (e.g. by their personal assistants), duplicates, and those emails which contained only forwarded content were removed from the corpus, to ensure that analysis was only dealing with the language of the four traders. This 'Trader Sent Corpus', was designed to meet as many as possible of Kredens' (2002:406) criteria that comparative language data should meet for it to be 'maximally homogenous'. The authors are as similar to each other as possible in terms of social characteristics; they are all traders, that is, they dealt with buying and selling energy and commodities in Enron's online market place, they are all male, and are all of working age. Further, the mode of communication is the same for all authors, and because they are doing the same job, it may be assumed that the subject matter of the emails will be generally similar in most cases. The argument is that if linguistic variation is found between these writers then 'it would be as a matter of course proportionately greater between speakers of dissimilar characteristics' (Kredens' 2002:406).

For many text-types involved in authorship casework there may be very little data available to the analyst. However, growingly cases of authorship assignment are involving digital texts such as emails, instant messaging and text messaging (Coulthard, Grant and Kredens 2011:538). Most of these text types, particularly emails, are digitally stored and easily retrievable, and the forensic linguist may have at their disposal large quantities of known documents for comparison. Thus, the methods applied and results obtained in this corpus-based empirical research may be useful and applicable to real authorship cases involving email.

3.2 The 'Enron Sent Email Author Reference Corpus'

After the author-distinctive linguistic features are identified within the Trader Sent Corpus, the second stage of analysis examines the frequency of these forms in the much larger 'Enron Sent Email Author Reference Corpus' (ESEARC), comprising another 126

Enron employees. Woolls (2012) extracted all of the various different sent folders ('sent', 'sent_items', or '_sent_items) for each of the 126 authors, giving a total of 40,236 emails and 1,669,197 words. Thus, in relation to the Trader Sent Corpus, ESEARC contains 32 times more authors, 15 times more emails and 19 times as many words.

The use of large-scale reference corpora is considered an important development in forensic authorship analysis, particularly in the exploration of idiolect and for measuring the evidentiary weight and diagnostic power of stylistic variables identified (Cotterill 2010; Solan 2012:367). 'Reference' corpora are referred to as such because they aim to offer 'normative data' (Kredens 2002:435), against which the rarity and expectancy of a particular linguistic feature can be measured. Related to this, Grant (2010:515) introduces the measurement of 'population-level distinctiveness', that is, if one person's 'style can be said to be distinctive, unusual or even unique against a reference population'. This forms the research question for the second part of the analysis in this study; what is the population-level distinctiveness of the linguistic features identified in the Trader Sent Corpus when compared with the ESEARC?

The way that this population-level distinctiveness is measured in this analysis is by calculating 'ratios of likelihood', or 'likelihood ratios'. Likelihood ratios are commonly used in medicine, where they are considered to be 'one of the best measures of diagnostic accuracy' (McGee 2002). Coulthard (2010:482) discusses the use of likelihood ratios in relation to forensic linguistic consultation, insofar as calculating the likelihood that a text would be in a particular form if the writer in question had or had not written it. This hinges on comparing the number of times the author uses a particular variant as a proportion of all occurrences of that variable, with the number of times that this variant was found 'in a representative sample of [text] messages produced by the general population' (Coulthard 2010:482). In the analysis here, when the distinctiveness of features identified in the Trader Sent Corpus is tested against ESEARC, the likelihood ratio is calculated by:

$$\frac{probability\ of\ greeting\ or\ farewell\ form\ occurring\ in\ an\ email\ written\ by\ a\ particular\ trader}{probability\ of\ greeting\ or\ farewell\ form\ occuring\ in\ an\ email\ in\ ESEARC}$$

So, for example, if a certain greeting form was used in 10% of a particular trader's emails, but was only found in 5% of emails in ESEARC, this would produce a likelihood ratio of 2 (i.e. 10%/5%). That is, it is twice as likely that this greeting form would occur if the email was written by the trader in question than if it were written by another author in the ESEARC. Similarly, if a form was found in 50% of a particular trader's emails, but only 0.1% of emails in ESEARC, that would produce a much higher likelihood ratio of 500 (50%/0.1%), meaning that it is 500 times more likely that an email would contain this form if it were written by that trader than if it were written by another author in ESEARC. Likelihood ratios can range from 0 to infinity, and as shown by the two hypothetical examples here, the larger the number the more distinctive the variant is of the trader or author in question.

One issue with this method which Coulthard (2010:483) acknowledges, is 'how does one establish what is a relevant population' of writers for comparison purposes? The ESEARC is not representative of the general population of writers; rather, the key word here is *relevant* population. Kredens (2002:435) argues that 'a potentially efficacious method should thus use reliable reference data, characterised by biological, social and interactional variables identical with those of samples A and B'. The ESEARC meets some of the criteria suggested by Kredens insofar as many variables remain constant between the emails in the Trader Sent Corpus and the ESEARC, and it is an effective reference corpus for ensuring we are comparing emails written by Enron employees with emails written by Enron employees.

Indeed, this comparison of like-for-like documents is encouraged by Butters (2012:354) in his discussion of standards and best practices for authorship analysis.

4 Method: Analysing email greetings and farewells

In terms of digital communication genres, email is one of the oldest and as a result it enjoys substantial attention in terms of genre features and conventions. This is particularly true of the two conventions which are the focus of this analysis – greetings and farewells – the linguistic forms and functions of which are now well-reported. Gains (1999) for example, compares the email openings and closings across commercial and academic emails, finding that the two sources of email data share conventions of openings, namely the use of no greeting, *hi*/*hello* greetings and *Dear* greetings. He also identifies a correlation in commercial emails of *thank you* closings in messages making requests, and closings using the sender's name only in information-giving messages. Lan (2000:55) compares email language of native English speakers and non-natives, finding generally that non-native speakers 'are more likely to stick to the rules of formal writing', particularly in the prominence of the formal '*dear...*' greetings. Waldvogel (2007) focuses on greetings and closings in two comparative New Zealand workplaces, an educational organisation and a manufacturing plant. She found that in the educational setting greetings and closings were not widely used, while in the manufacturing plant they were used extensively, including naming greetings and closings, and formulae such as *hi* and *dear* and *thanks*, *cheers* and *regards*. She argues that the overriding factor in accounting for these differences is workplace culture; namely that low morale in the educational organisation has resulted in socially distant communication styles while the greetings and farewells used in the manufacturing plant reflected open and positive relationships between staff. Biber and Conrad (2009:189) compare the salutation and signature patterns in personal emails to family and friends with professional emails to colleagues and strangers, arguing that differences in role relationships between participants account for different patterns of use within these conventional elements. For example, emails between colleagues frequently dispose of salutations, while emails to family and friends display *Hi+Name* greetings and emails to strangers use formal *dear* with *Title+Last name* forms. With signatures, emails between friends and family most commonly end with the sender's name only, emails to colleagues include a formula such as *best wishes* and *thanks*, while emails to strangers make use of formal signatures such as *sincerely*. Similarly, Bou-Franch (2011:1783) analyses variation within the opening and closing of emails in Spanish writers in relation to institutional power of participants and interactional position of an email in the overall conversation. She found that emails sent down the power hierarchy contained "the least dense opening and closing sequences, the smallest number of greetings, self-identifications and thanking moves" when compared to those sent up the power hierarchy to dominant participants. Also, as the email conversation unfolds, openings and closings displayed less elaboration and more intimacy.

Overall, then, despite these linguistic elements of emails being inherently 'conventional', there is often a remarkable range of variation within them (Crystal 2001:106). This variability presents greetings and farewells as of use to the authorship analyst, particularly given that 'the usage of greeting and/or farewell text [...] may be as habitual as vocabulary or syntax' (Corney, Anderson, Mohay and de Vel 2001:4). Therefore, this study can build upon and contribute to this body of literature relating to email greetings and farewells, but with a particular focus on author-distinctive patterns of use.

In identifying greetings and farewells, a greeting was considered to be anything that preceded the main content body of the email, and a farewell is the text which follows the main body of the email, forming the very last part of the message, excluding any signatures

automatically generated by email client software (Figure 1). It is useful to consider the analysis of greetings and farewells within the sociolinguistic construct of the linguistic variable, 'a linguistic item which has identifiable variants' (Wardhaugh 2006:143). The different forms being used by the traders constitute 'variants' of the greeting and farewell 'variables'.

The coding of the variants in the Trader Sent Corpus and ESEARC was performed both qualitatively and computationally, and a very fine-grained approach was adopted, taking into account differences in punctuation, line spacing and capitalisation. That is, if two greeting forms differ only in the capitalisation of their first letter, then they are considered different variants. The advantage of focusing on generic elements of emails is that the authors necessarily use one variant or another in their writing, even if the variant is to omit the greeting or farewell altogether, and so every email provides an occurrence of the variable.

[FIGURE 1 NEAR HERE]
Figure 1: Example of a greeting and farewell in the analysis

5 Analysis: Distinctive variation within email conventions in the Trader Sent Corpus

5.1 Greetings

The first stage of analysis aims at identifying greeting and farewell variants which distinguish between the emails of the four traders in the Trader Sent Corpus. Across the 2,622 emails there are a total of 73 greeting forms used by the four traders, giving an average in the corpus of a different greeting form for every 36 emails. The traders display similar levels of variation to each other, each using between 27 and 35 different greeting forms (Table 2). In this analysis, only those forms which are used more than once by an author are included, to maintain a level of intra-author consistency.

Table 2: Number of emails per different greeting across the four authors in the Trader Sent Corpus.

	Arnold (632 emails)	**Germany** (1339 emails)	**Lavorato** (405 emails)	**Zipper** (246 emails)
Number of greeting forms	23	41	12	7
Emails per greeting form	27	33	34	35

Within the variants that have been retained for discussion here, clear patterns of use emerge, many of which may be considered author distinctive, or at least give an indication of authorship, within the Trader Sent Corpus (Table 3). They can be split into three main types of greeting, which align with findings of many of the studies outlined above:

- No greeting
- Naming greetings
- *Hey/hi/hello* greetings

5.1.1 No greeting

Overwhelmingly the most common choice by all of the traders is to omit the greeting altogether. This is not surprising given that studies of email in professional and personal

Table 3: Greeting forms and their use across the traders

Greeting	Arnold (n=632)	Germany (n=1339)	Lavorato (n=405)	Zipper (n=246)
No greeting	476 (76%)	1190 (89%)	255 (63%)	205 (83%)
Name:	106 (17%)		1 (0.2%)	
Name, (next word capitalised)			11 (3%)	34 (14%)
Name, (next word not capitalised)		66 (5%)	4 (1%)	1 (0.4%)
Name.			5 (1.2%)	
Name (next word capitalised)			124 (31%)	2 (0.8%)
Name (next word not capitalised)	1 (0.2%)		1 (0.2%)	2 (0.8%)
Name –		1 (0.1%)	1 (0.2%)	
Total naming greetings	107 (16%)	67 (5%)	147 (36%)	39 (16%)
Hey:	17 (2.3%)			
Hey, (next word caps)	3 (0.5%)			
Hey, (next word not caps)		10 (0.7%)		
Hey.		2 (0.1%)		
Hey!!		4 (0.3%)		
Hey + group*		11 (0.8%)		
Hey + *buddy*		6 (0.4%)		
Hey + Name,	2 (0.3%)	9 (0.7%)	1 (0.2%)	
Hey + *woman*		2 (0.2%)		
Hi:	2 (0.3%)			
Hi.		2 (0.1%)		
Hi + group*		21 (1.6%)		
Hi + First Name,		3 (0.2%)		
Hello:	7 (1%)			
Hello,	3 (0.5%)			
Total *hey/hi/hello* greetings	34 (5%)	70 (5%)	1 (0.2%)	0

*Subsumes *guys, gang, team, dudes, campers*

domains have traditionally found that writers most commonly dispense with opening routines (Gains 1999:85; Waldvogel 2007:462; Biber and Conrad 2009:189). Crystal (2001:100) states, 'between people who know each other, greetingless messages are usually promptly sent responses [...] for which an introductory greeting is inappropriate'. This immediately underlines the important effect that communicative context has on greeting and farewell choices. The relationship between sender and recipient, the topic and purpose of the email, the speed of correspondence, and whether the email is initiating conversation or is a reply, will all have an effect on the greetings and farewells chosen by the authors. Although this sociolinguistic information has not been systematically analysed here, such influences on linguistic choices are considered throughout the analysis.

5.1.2 Naming greetings

Besides having no greeting, the most frequently occurring greeting variants are those which include only the recipient name(s). Within this type of name greeting, author-distinctive or author-typical patterns can be identified for all four traders. In the first instance, Table 3 shows that Lavorato uses naming greetings far more frequently than the other four traders (36% of his emails). In particular, his greeting style is characterised by his choice not to use any punctuation after his recipient's name (31%) (Example 1). Although Zipper also makes use of this form, he does so far less frequently than Lavorato (<1%). Lavorato is also the only one of the four traders to use a full stop after the name of his recipient (1.2%) (Example 2). In contrast, a greeting in which the recipient's name is followed directly by a colon is distinctive of Arnold's emails (17%) (Example 3), with only one occurrence in the writing of the other three traders. Further, both Germany (5%) and Zipper (14.4%) make relatively frequent use of the comma after the recipient's name. However, their usage is different insofar as Zipper almost always capitalises the next word of the email following the greeting (there is one exception), while Germany never does (Examples 4 and 5). Again, although Lavorato uses both of these forms, he does so with far less frequency and consistency than Zipper and Germany.

(1) **Kim**
Have you called Nigel and told him about the interviews. (Lavorato83)

(2) **Milly**.
If you're going to sell the warrants yourself. (It sounds like you should if your getting $2 in time value) let me know and I'll call off the equity group. (Lavorato339)

(3) **Gary:**
Just checking to see if the Trader's Roundtable includes us gas boys. I would certainly be interested in attending.
John (Arnold19)

(4) **Ras,**
This looks fine. Please distribute to Accenture.
Thanks,
AZ (Zipper35)

(5) **Kim,**
please send Victor and myself an amtel message at 11:00am AND at 1:30pm on Feb 1 to attend the meeting below.
Please include the room and time on the amtel.
I think it will be a busy day.
Thanks (Germany599)

5.1.3 *Hey/hi/hello* greetings

The second major type of greeting form used by the four traders comprises those greetings which include the use of either *hey, hi* or *hello* followed by some punctuation mark or form of address. Generally these forms are of lower frequency than those using the recipient's name. However, there are still several author-distinctive patterns within this greeting type, particular with regards to Arnold and Germany's greeting styles.

In the same way as the use of a colon distinguishes Arnold's naming greetings from those of the other traders, he is also the only of the four traders to use a colon after *hey, hi* and *hello* in greetings (3.6%) (Example 6). Arnold is also the only author to use a comma after *hey* and *hello* with the following word capitalised (1%) (Example 7). Germany uses a much wider range of *hey/hi/hello* greeting variants than the other traders. First, he makes use of a very similar form to Arnold's *hey/hello*+comma, but does not capitalise the next word (0.7%) (Example 8), and is the only author to do this. In addition, Germany has a further two greetings that are only found in his emails; he is the only trader to follow *hey* and *hi* with a full stop (0.2%) (Example 9), which can be compared with Lavorato's use of a full stop after recipients' names, discussed above. Germany is also the only author to follow *hey* with (double) exclamation mark(s) (0.3%) (Example 10). Besides these, Germany's greeting style is also distinguished by addressing a group of recipients after *hey* and *hi*, using *guys, gang, team, dudes* or *campers* (0.8%) (Example 11). He is also the only of the four traders to greet his recipient with *hi* followed by their first name (0.7%) (Example 12). Finally, in a small number of emails, he greets his reader with *hey buddy* (0.4%) (Example 13) and *hey woman* (0.2%) (Example 14).

(6) **Hey:**
I hope you know we were just kidding with you yesterday.
We don't get many strangers on the floor so we have to harass them when we do.
I think I am a couple of the "we are not" attributes in your book.
Is that going to cause me any problems going forward?
John (Arnold31)

(7) **Hello,**
Just checking to see if things are progressing as scheduled.
Thanks,
John (Arnold181)

(8) **Hey,**
you're back. Did you have fun?
jerry and I are taking a motorcycle trip next week. (Germany867)

(9) **Hi.** I will not be able to attend because I'm dislocated my ankle and I'm crutchin it now (Germany499)

(10) **Hey!!** Where are you? I'm about to go get my Dad a chocolate malt. Today is his long day at the hospital. He's probably still there (Germany572)

(11) **Hey guys.**
I have a ton of emails to plow thru with CES buy look what I found.
The CES storage proxy schedule is at the bottom. (Germany191)

(12) **Hi Steph,**
don't know if this question should go to you. If not, please forward it on to the right person and let me know who that is.
The reservation volumes on the seasonal contracts do not look correct.
The demand charge looks OK, its the demand volume which is not calculating correctly [...] (Germany1292)

(13) **Hey buddy.**
I would love to take a tour.
I'm tied up with a family illness for the next couple of weeks so don't wait on me. Great Idea though (Germany1218)

(14) **Hey woman,**
how much vacation do I have left BEFORE i fill out a timesheet for the 2nd part of August? (Germany1162)

The difference in the choice of greetings is undoubtedly determined, at least to some extent, by the differences in context. The vocative greetings may be used more frequently in emails which 'attempt to get someone to do something' (Holmes 2001:259), which appears to be the case in Examples 1-5 above. In contrast, greetings such as *hey*, *hi* and *hello* may be considered more informal, and along with *buddy* and *woman*, indicative of a more intimate social relationship between interlocutors, found in emails with a social purpose rather than a transactional one, such as in the majority of Examples 6-14.

5.2 Farewells

There are 73 different farewell variants found in the Trader Sent Corpus, the same as the amount found for greetings. However, the variation between the authors is quite different to that of greetings. Zipper was the least variable in terms of his greetings, yet he is by far the most variable in his use of farewells, and Germany is by far the least variable in his use of greetings, with only 27 different farewell forms across over 1300 emails (Table 4). This shows that email writers can be more variable in their linguistic selections within one email convention than they may be in others. Again, only those forms which are used more than once by a trader are included here.

Table 4: Number of emails per different farewell across the four authors in the Trader Sent Corpus (compared with greetings)

	Arnold (632 emails)	**Germany** (1339 emails)	**Lavorato** (405 emails)	**Zipper** (246 emails)
Number of farewell forms	32	27	19	30
Emails per farewell form	20	50	21	8
Number of greeting forms	23	41	12	7
Emails per greeting form	27	33	34	35

Having no farewell is the most common choice by all of the traders, possibly due to the same factors relating to omitted greetings. In total, 72% (1879/2622) of all of the emails in the Trader Sent Corpus have no greeting and no farewell; 68% (432/632) of Arnold's emails, 80% (1068/1339) of Germany's, 52% (210/405) of Lavorato's and 69% (169/246) of Zipper's have no greeting and no farewell. Thus, this is the unmarked choice for all of the four authors. Again, though, in the farewells that are used, there are observable patterns of distinctive and individuating language use (Table 5). As with the greetings, the farewells used across the authors can be categorised into major recurring types:

- Naming farewells
- *Thanks* farewells
- *Regards/love/later* farewells

Table 5: Farewell forms and their use across the traders

Farewell form	**Arnold (n=632)**	**Germany (n=1339)**	**Lavorato (n=405)**	**Zipper (n=246)**
No farewell	470 (74%)	1189 (89%)	314 (78%)	172 (70%)
First Name	78 (12%)	2 (0.2%)	32 (8.2%)	28 (11.4%)
First Name.	3 (0.4%)		5 (1%)	
Last Name			26 (6%)	
Comma + First Name	6 (1%)			1 (0.4%)
First Name + Last Name	2 (0.4%)	2 (0.1%)	4 (1%)	4 (2%)
Name + Company details	4 (0.6%)			1 (0.4%)
One initial	2 (0.3%)			
Two initials (capitalised)				7 (3%)
Two initials (not capitalised)		1 (0.1%)		1 (0.4%)
Total naming farewells	95(15%)	5 (0.4%)	67 (17%)	42 (17%)
Thanks	3 (0.5%)	114 (9%)	6 (1%)	13 (5%)
Thx (+ Name)	6 (0.9%)			
Thanks, First Name	44 (7%)			3 (1.2%)
Thanks. First Name				1 (0.4%)
Thanks First Name	2 (0.4%)	2 (0.2%)	8 (1.5%)	4 (1.4%)
Thanks, Full name	1 (0.2%)	1 (0.1%)	3 (0.7%)	1 (0.4%)
Thanks initials (not capitalised)		7 (0.5%)		
Thanks, initials (capitalised)				6 (2.4%)
Total *thanks* farewells	56 (9%)	124 (9%)	17 (4%)	28 (11%)
Regards + First Name			2 (0.4%)	3 (1%)
Regards + Full name			4 (1%)	
Love you + First Name	2 (0.3%)			
Love you (and the kids) mucho, KPD				2 (0.8%)
Later		4 (0.3%)		
Total *regards/love/later* farewells	2 (0.3%)	4 (0.3%)	6 (1%)	5 (2%)

5.2.1 Naming farewells

This category includes all of the farewells which comprise the first name, last name or initials of the sender. Overall, as was the case with the greetings, Germany uses naming farewells (5%) far less frequently than the other three traders. The most common of the naming farewells is first name only, which is common in emails between colleagues (Gains 1998:86; Lan 2000:26; Biber and Conrad 2009:190). Given the popularity of this form, it does not serve to distinguish between the four traders. However, the use of *last* name only (more specifically a clipped version of his last name) as a farewell is distinctive of Lavorato's emails, as he is the only of the four authors to do this (6%) (Example 15). The only other individuating naming farewells both involve the use of initials. First, Arnold is the only author to sign off with just one of his initials (0.3%) (Example 16). Second, while Germany

and Zipper both sign off with their initials in lower case (though only once each), Zipper is the only trader to sign off with his initials capitalised (3%) (Example 17).

(15) You might want to file this. **Lavo** (Lavorato290)

(16) I have a membership to the Body Shop downstairs.
Can you cancel that please?
J (Arnold129)

(17) Congratulations !
Now let's focus on closing the deal and getting someone else signed up.
Well done all.
AZ (Zipper20)

5.2.2 *Thanks* farewells

Less common than naming farewells such as these, but which are used by all four traders, are sign-offs which include *thanks* either standing alone or followed by the sender's name. With instances in which the recipient's name does follow *thanks*, it has been counted as a *thanks* farewell variant as opposed to a naming variant. Despite the popularity of *thanks* farewells across the Trader Sent Corpus, author-distinctive forms still emerge, with particular variants being distinctive of Arnold's, Germany's and Zipper's email style.

Arnold's email style is characterised by the use of the farewell form which combines *thanks* with his name separated by a comma (7%) (Example 18). Although Zipper also uses this form, he does so in only 1.2% of his emails; thus, this is another case in which a variant is not strictly individuating of a single trader's emails, but is used far more consistently by one of the authors. Further, Arnold is the only author to use the abbreviation *thx* either as a standalone farewell or followed by his first name (0.9%) (Example 19).

(18) John: Please call Lavorato's secretary, Kim, and schedule a time to talk with John ASAP.
Thanks,
John (Arnold37)

(19) Can you swap me with Fletch.
Try to make all of my interviews as late as possible.
Thx
John (Arnold57)

The simple stand alone sign-off of *thanks* with no punctuation or name is used by all four of the traders. However, it occurs in Germany's emails far more consistently than the others'; it is found in 9% of his emails compared to 0.5%, 1% and 5% of the others (Example 20). Furthermore, Germany and Zipper both use very similar forms of *thanks* followed by their initials. The difference here is that Zipper capitalises his initials (2.4%) (Example 21), while Germany does not (0.5%) (Example 22), and they are the only authors in the Trader Sent Corpus to use these respective variants.

(20) As usual, you are awesome!

Let Steve know what the fuel waivers are when you get them.
Thanks (Germany65)

(21) Ras, This looks fine.
Please distribute to Accenture.
Thanks,
AZ (Zipper35)

(22) We should have 4682 of Egan coming into TGT for the 1st and 4679 fo the 2nd -31st. Would you guys check that please.
thanks
cg (Germany817).

5.2.3 *Regards/love/later* farewells

Finally, there are a number of low-frequency farewell variants that do not belong to naming farewells or *thanks* forms, but that are only used by one of the four authors. *Regards* followed by the full name of the sender is only used by Lavorato (Example 23). Similarly, *love you* appears in the farewells of both Arnold and Zipper, but whereas Arnold follows *love you* with his first name (Example 24), Zipper follows his with *mucho* (Example 25). He also signs off with the curious acronym *KPD*, which could be a nickname or pet name Zipper has with presumably his wife or partner, or it may be a humorous reference; an internet search suggests *KPD* as an acronym for *knobs per dollar*. Lastly, Germany is the only of the four traders to use *later*, which he either uses as a standalone farewell, or to precede *gator, buddy* or *dudes* (Example 26).

(23) As you know we were trying to do some funding deals prior to the end of the year.
We executed a total of $65 Million CAD with 4 different companies (TD, CIBC, Bank Paribas, and BMO).
Everyone who I cc'd this note to did a great job of helping us execute.
Regards
John Lavorato (Lavorato422)

(24) Thank you very,very much.
Love you,
John (Arnold237)

(25) Hi !
thanks for message.
I will try to call you on cell phone.
Love you mucho,
KPD (Zipper98)

(26) Hi Team!
I will be out of the office on Friday, March 31st.
Please call Scott Goodell at 713-853-7711 with any questions you may have.
Later (Germany256)

The authors' choice of farewell variants will be influenced by the function of the email and their recipient(s). For example, Gains (1998:86) suggests a link between requests in emails and *thank you* farewells, and this pattern emerges in Examples 18-20. In turn, as discussed above, vocative greetings may be linked to requests; as such, we get greeting/farewell combinations such as that in Examples 18 and 21. The result of this may be that, for some authors, vocative + request + thanks ..., are pre-conditioned and co-selected variants. In the same way, more informal emails with intimate interlocutors may contain greeting/farewell combinations such as *Hi + Love you,* as in Example 25. Further, the omission of a farewell form may be pre-determined by the omission of a greeting form. Thus, we arrive at a situation in which certain co-occurrences and co-selections become distinctive of particular authors, and the idea of such combinations of variants is revisited below.

Overall, this first stage of analysis has identified a number of observable patterns in greetings and farewells which can be used to distinguish between the traders. The findings here justify the fine-grained coding of formal differences in variants, given that, in many cases what separates two authors' usage is the inclusion or omission of a particular punctuation mark or the decision to capitalise words or not. Indeed, there is an argument that these types of orthographical choices are least likely to be consciously controlled by the author (Johnson 2012). A result of this fine-grained approach is that many of the distinctive and individuating forms are low-frequency, sometimes occurring in less than 1% of an author's emails. However, what remains true is that in the context of the Trader Sent Corpus, such forms are distinctive or individuating of particular writers.

6 Analysis: Population-level distinctiveness of features identified

The analysis above has identified a total of 19 greetings and 12 farewells that are distinctive or individuating of the four authors in the Trader Sent Corpus, and these are shown in column one of Tables 6 and 7 below. However, no comment can yet be made of the distinctiveness of these forms, or the authors' email-lects or idiolects in a wider population of writers. To that end, this second stage of analyses explores the frequency of these 31 forms in the Enron Sent Email Author Reference Corpus (ESEARC) of 126 authors, 40,236 emails and 1,669,197 words. The forms were found in the ESEARC computationally. Many of the greeting forms such as those including *hi, hello, hey* (and their collocates) and farewell forms such as *thanks* (and its collocates), *love, regards* and the senders' names were searched for directly in the corpus using *Wordsmith Tools* (Scott 2008). On the other hand, this approach could not be used to identify instances in which the authors in ESEARC used their recipients' names in their greetings. Therefore, when extracting the ESEARC data from the CMU set, Woolls (2012) produced a complete list of all of the first words appearing in the emails of the 126 authors, and so recipients' names as used in greetings were captured.

Finding the frequencies of occurrence of the forms in the ESEARC, and using likelihood ratios, allows us to measure their expectancy, rarity and distinctiveness in a relevant population of writers. Tables 6 and 7 present the results for the greetings and farewells respectively. The tables compare the frequency of occurrence of the variants in the emails of the traders of which they are initially distinctive, with their occurrence in the ESEARC. Undergoing this test, there are a number of features which lose their distinctiveness, while some remain remarkably distinctive of the trader who uses them.

6.1 Variants which lose their distinctiveness

The greeting *Hi.* was distinctive of Germany in the Trader Sent Corpus, but occurred in only 0.1% of his emails. This form was found in 34 of the 40,236 emails in ESEARC, or 0.1%. Thus, calculating the likelihood ratio for this variant by dividing these two probabilities by one another (0.1%/0.1%) gives a likelihood ratio of 1. The same is the case with the

Table 6: Population-level distinctiveness of greetings identified in the Trader Sent Corpus

Trader	Variant	Freq. in Trader Sent Corpus	(%)	Freq. in ESEARC (n=40,236)	(%)	No. of authors in ESEARC (n=126)	(%)	Likelihood ratio
Arnold (n=632)	*Hello*:	7	(1)	1	(0.002)	1	(0.8)	**500** (1/0.002)
	Hey:	17	(2.3)	4	(0.01)	2	(1.6)	**230** (2.3/0.01)
	Hello,	3	(0.5)	6	(0.02)	3	(2.4)	**25** (0.5/0.02)
	Hi:	2	(0.3)	7	(0.02)	1	(0.8)	**15** (0.3/0.02)
	Name:	106	(17)	668	(1.7)	20	(15.9)	**10** (17/1.7)
	Hey, (next word caps)	3	(0.5)	18	(0.1)	11	(8.7)	**5** (0.5/0.1)
Germany (n=1339)	*Hey* + *woman*	2	(0.2)	0	(0)	0	(0)	-
	Hey + *buddy*	6	(0.4)	6	(0.01)	2	(1.6)	**40** (0.4/0.01)
	Hey!!	4	(0.3)	5	(0.01)	5	(4)	**30** (0.3/0.01)
	Hi + group	21	(1.6)	28	(0.1)	6	(4.8)	**16** (1.6/0.1)
	Hey.	2	(0.1)	4	(0.01)	4	(3.2)	**10** (0.1/0.01)
	Hey + group	11	(0.8)	27	(0.1)	12	(9.5)	**8** (0.8/0.1)
	Hey, (next word not caps)	10	(0.7)	48	(0.12)	14	(11.1)	**6** (0.7/0.12)
	Name, (next word not caps)	66	(5)	1174	(3)	48	(38.1)	**2** (5/3)
	Hi.	2	(0.1)	34	(0.1)	3	(2.4)	**1** (0.1/0.1)
	Hi + First Name,	3	(0.2)	581	(1.4)	44	(34.9)	**0.14** (0.2/1.4)
Lavorato (n=405)	Name.	5	(1.2)	15	(0.04)	10	(7.9)	**30** (1.2/0.04)
	Name (next word caps)	124	(31)	2026	(5)	101	(80.2)	**6** (31/5)
Zipper (n=246)	Name, (next word caps)	34	(14)	2105	(5)	82	(65.1)	**3** (14/5)

Table 7: Population-level distinctiveness of farewells identified in the Trader Sent Corpus

Trader	Variant	Freq. in Trader Sent Corpus	(%)	Freq. in ESEARC (n=40,236)	(%)	No. of authors in ESEARC (n=126)	(%)	Likelihood ratio
Arnold (n=632)	*Love you* + First Name	2	(0.3)	3	(0.01)	1	(0.8)	**30** (0.3/0.01)
	Thx (+ Name)	6	(0.9)	48	(0.1)	13	(10.3)	**9** (0.9/0.1)
	Thanks, First Name	44	(7)	2938	(7)	77	(61.1)	**1** (7/7)
Germany (n=1339)	*Thanks*	114	(9)	557	(1.4)	60	(47.6)	**6** (9/1.4)
	Thanks initials (not caps)	7	(0.5)	66	(0.2)	8	(6.3)	**3** (0.5/0.2)
	Later	4	(0.3)	32	(0.1)	15	(11.9)	**3** (0.3/0.1)
Lavorato (n=405)	Last Name	26	(6)	49	(0.1)	9	(7.1)	**60** (6/0.1)
	Regards + Full name	4	(1)	101	(0.3)	16	(12.7)	**3** (1/0.3)
Zipper (n=246)	*Love you (and the kids) mucho*	2	(0.8)	0	(0)	0	(0)	**-**
	Thanks, initials (caps)	6	(2.4)	569	(1.4)	11	(8.7)	**2** (2.4/1.4)
	Two initials (caps)	7	(3)	1517	(3.8)	24	(19)	**0.8** (3/3.8)

farewell *Thanks* + First Name distinctive of Arnold. This was found quite commonly in his emails, being used in 7% in total. However, it is very common in ESEARC, occurring in 2938 emails, or 7%. Thus, again, the likelihood for this variant is 1 (7%/7%). In other words, based on these population findings, an email is just as likely to contain these greeting and farewell variants if it were written by a member of a relevant population as if it were written by Germany or Arnold. That is, these variants have lost their distinctiveness altogether.

Further, the greeting *Hi*+First Name is characteristic of Germany in the Trader Sent Corpus. However, while it only occurs in 0.2% of his emails, it is found in 1.4% of ESEARC emails, giving a likelihood ratio of 0.14 (0.2%/1.4%). Similarly, the farewell using two capitalised initials was typical of Zipper in the trader corpus, being used in 3% of his emails. At the same time, however, this farewell is found in 3.8% of ESEARC emails, giving a likelihood ratio of 0.8 (3%/3.8%). Given that these two forms have produced likelihood ratios of less than 1 at this population level, they are both *less* likely to be found in an email written by Germany or Zipper than in an email written by another member of the relevant population.

6.2 Variants which remain distinctive of a particular trader

6.2.1 Greetings

In total, there are 17 greeting forms which remain distinctive of the trader who uses them, that is, they have a likelihood ratio of more than 1. Of course, some are more distinctive than others. The greeting *Hey woman* is not found at all in the ESEARC, and so remains entirely individuating of Germany's emails. Besides this variant, there are a range of high likelihood ratios produced for greetings and farewells. Generally, the higher likelihood ratios belong to greetings, that is, greetings identified in the Trader Sent Corpus have higher population-level distinctiveness than the farewells. Space does not permit a full description of each form here, so a number of examples have been selected for discussion. Arnold has the most distinctive greeting variants (Table 6). His use of *Hey:* and *Hello:* with a colon are rare in ESEARC. *Hey:* is found in 2.3% of his emails, but only 0.01% of the emails in ESEARC, giving a likelihood ratio of 230 (2.3%/0.01%). Similarly, although *Hello:* is found in only 1% of Arnold's emails, it is found only once (0.002%) in the ESEARC, giving a likelihood ratio of 500 (1%/0.002%). Therefore, it is 230 times more likely that the greeting *Hey:* would occur if an email were written by Arnold than if it were written by a member of a relevant population, and it is 500 times more likely that *Hello:* would occur.

Germany and Lavorato also have some relatively distinctive greetings (Table 6). Germany uses *Hey!!* in 0.3% of his emails, and it only appears in 0.01% of ESEARC emails, giving a likelihood ratio of 30 (0.3%/0.01%). Similarly, he uses *Hey buddy* in 0.4% of his emails, compared with the 0.01% of emails it is found in ESEARC, giving a likelihood ratio of 40 (0.4%/0.01%). Thus, it is 30 times more likely that *Hey!!* will occur in an email produced by Germany than an email produced by another author, and 40 times more likely that *Hey Buddy* will be found. For Lavorato, the Name+full stop form is used in 1.2% of his emails, and is only found in 0.04% of ESEARC emails, and so has a likelihood ratio of 30 (1.2%/0.04%). While these variants are not as strongly distinctive as Arnold's colon use, they do have some level of diagnostic power at population level.

6.2.2 Farewells

In total there are eight farewell variants which remain distinctive of the trader who uses them. Generally with the farewells, though, all of the likelihood ratios are relatively low in comparison with the greetings (Table 7). One exception to this is the farewell *Love you*

mucho used by Zipper, which is not found at all in ESEARC, and so remains individuating of his emails.

The farewell *Love you*+First Name, is distinctive of Arnold in the Trader Sent Corpus, being used in 0.3% of his emails. This form is only found in 0.01% of ESEARC emails, and so has a likelihood ratio of 30 (0.3%/0.01%). Similarly, the use of the sender's last name, or a clipped form of the last name, is distinctive of Lavorato in the Trader Sent Corpus, as he uses it in 6% of his emails. This farewell variant appears in 0.1% of ESEARC emails, producing likelihood ratio of 60 (6%/0.1%). Therefore, the farewells *Love you*+First Name and the Last Name form are 30 and 60 times more likely to occur written by Arnold and Lavorato respectively than if they were written by another author in a relevant population.

6.3 Combinations of variants

These population-level results show how distinctive a single greeting or farewell form can be. As Coulthard (2010:482) comments, a major advantage of using the likelihood ratio method is that it can 'take account of co-selections by combining several independent ratios together to produce a composite likelihood ratio [...] by simple multiplication'. Thus, this method can be used to calculate the likelihood of co-occurrence of any greeting and farewell variants in the same email. For example, in two of his emails, Arnold uses the Name: greeting together with the *Thx* (+Name) farewell (Example 27). Table 6 and 7 show that these forms have likelihood ratios in ESEARC of 10 and 9 respectively, producing a composite likelihood ratio of 90 (10x9). That is, it is 90 times more likely that this combination of greeting and farewell would occur if the email was produced by Arnold than if it were produced by a member of the relevant population. Similarly, in four emails, Germany uses the *Hi*+group greeting with the standalone *Thanks* farewell (Example 28). These two variants have a likelihood ratio of 16 and 6 respectively (Table 6 and 7), giving a combined likelihood ratio of 96. Finally, in ten of his emails, Lavorato combines the greeting of the recipient's name with no punctuation and the next word capitalised, with the farewell of just his last name (Example 29). This greeting has a likelihood ratio of 6 (Table 6) while the farewell has a likelihood ratio of 60 (Table 7), giving a combined ratio of 360. Thus, this email is 360 times more likely to include this greeting and farewell combination if Lavorato had written it than if it had been written by another member of a socially-similar population of writers.

(27) **Dave:**
A couple things about the limit orders:

1: When a customer opens up the limit order box, I think the time open should default to 12 hours. We want the orders open as long as possible.
[...]
4. Is there a way to modify a limit order, such as changing the price, without canceling it and resubmitting a new one? If not, this would be a valuable feature.
Thx,
John (Arnold93)

(28) **Hi gang.**
I want to meet tomorrow regarding the CES storage deal. Around 1:30PM? Let me know.
thanks (Germany101)

(29) **Dave**
We need to think about trying to get more VAR for the gas floor before all these other commodities go to the Board and scare the hell out of them.
I have several reasons for this so lets discuss.
Lavo (Lavorato162)

The findings from these analyses potentially have both theoretical and methodological implications. First, Grant (2010:515) comments that investigating population-level distinctiveness 'may have more profound implications for theoretical discussions of idiolect'. In Section 2 above, the notions of idiolectal style, discoursolect and genre-lects were introduced as more accessible, empirically analysable, and useful approaches to investigating an individual's idiolect. The results here are promising in that they suggest that an individual may have a distinctive or even unique email-lect or style. Even with as few as two variables – greeting and farewells – it has been shown that greeting and farewell choices, and combinations of these, can remain distinctive of individual writers when tested against a reference corpus of a relevant population of writers. With email, there is potential to expand analysis to incorporate additional conventions of the genre to include in composite likelihood ratios, including requests for follow up (e.g. *please let me know*), date formats, expression of attaching files to emails (e.g. *find attached* vs *attached is*) and non-standard punctuation and spelling . A hypothesis for future investigation is that, as the number of variables considered increases, so too does the probability of the combinations of variants within these genre conventions being unique to or distinctive of individual writers. In turn, this will provide further evidence to support the notion of author-unique email-lects.

Linked to this, it may be a productive methodological starting point for the forensic linguist to approach questions of authorship from a genre-focused perspective, using as their entry into the analysis the linguistic choices the author(s) in question makes within the conventions of the genre in which they are writing. A further important point to note from this analysis is the potential diagnostic power of low frequency variants. Butters (2012:356) questions the validity of low-frequency features and argues that a set of standards should 'suggest criteria for evaluating how meaningful an aggregated small number of variables can truly be'. While his comments relate to a case quite different to this analysis, the results here have shown that that, within an 87,000 word Trader Sent Corpus, low-frequency variables can be used to distinguish between the four authors. Moreover, when the analysis is expanded to the ESEARC, it is not always the most high-frequency of these distinctive variants that maintain their distinctiveness in a wider population of writers. For example, the greeting of the recipient's name with no punctuation (Example 29) is distinctive of Lavorato, found in 31% (n=124) of his emails. Within the Trader Sent Corpus, then, this form may be considered particularly powerful and reliable in distinguishing Lavorato's emails from the other three traders. However, in the ESEARC, this form is used by 101 authors, a total of 2026 times, producing a likelihood ratio of an unremarkable 6. Thus, the usefulness of a form such as this is restricted to cases which involve a low number of candidate authors. In contrast, Arnold's distinctive greeting *Hello:* was only found in 1% (n=7) of his emails, and so may be considered not as powerful or reliable as Lavorato's greeting in the first instance. However, *Hello:* is found only once in ESEARC, giving a very high ratio of likelihoods of 500. This relates to the linguistic notion of *markedness*, first introduced in terms of phonology and later extended to grammar and semantics (e.g. Jakobson 1956; Lyons 1968:79). In turn, it has been embraced by stylisticians as referring to 'any features or patterns which are prominent, unusual or statistically deviant in some way' (Wales 2001:244). Because omitting the greeting and farewell is by far the most common choice of authors, it is the unmarked for all of them (Section 5.2), any presence of a greeting or

farewell is statistically departing from this norm, and the lower the frequency, the further the deviation. Turell (2010:218–220) argues that the proposal that 'the marked form conveys more precise, specific and additional information than the unmarked form' is a most useful conceptualisation in the context of forensic authorship analysis. Therefore, offering an alternative to Butters' (2012:356) argument, it can be suggested that the less frequently an author uses a particular form, the more marked it is, and the more precise and specific information it can give about an author's style.

7 Conclusion

The analysis of the Trader Sent Corpus identified 19 greeting and 12 farewells which were either entirely individuating of one of the traders or were shared between two or more traders, but used far more consistently by one. The greetings identified belonged to three main types, no greeting, naming greetings and *hey/hi/hello* greetings. Particularly useful in distinguishing between the authors were Arnold's choice to follow the recipient's name, *hey, hi* and *hello* with a colon, Germany's greeting of multiple participants, his use of *buddy/woman,* and his use of the recipient's name or *hey* followed by a comma with the next word not capitalised. Lavorato was distinguished by his greeting of the recipient using their name with no punctuation or a full stop, while Zipper's emails were characterised by recipient's name followed by a comma with the next word capitalised. In terms of farewells, the main types were no farewells, naming farewells, *thanks* farewells and r*egards/love/later* farewells, with *Thanks* + his first name, and *thx* being distinctive of Arnold, stand alone *thanks* characterising Germany's emails, a clipped form of last name being typical of Lavorato, and capitalised initials being distinctive of Zipper.

The second stage of analysis introduced the 'Enron Sent Email Author Reference Corpus' (ESEARC) comprising 126 authors, 40,236 emails and 1,669,197 words. Using likelihood ratios, the expectancy, rarity and distinctiveness of the 31 greeting and farewell forms identified as useful in the Trader Sent Corpus were tested against this reference corpus of a relevant population of email authors. Four variants lost their distinctiveness altogether. Overall, however, especially for greetings, the results were encouraging, with some forms being 60, 230 and 500 times more likely to appear in an email written by the trader in question than another writer in the relevant population, namely Lavorato's last name only farewell, and Arnold's *Hey:* and *Hello:* greetings. The results of these tests highlighted the importance of low-frequency features in distinguishing between authors, as it was often the less frequently occurring variants which were the most distinctive at population level. Finally, it was shown how easily likelihood ratios for greetings and farewells can be multiplied to measure the author-distinctiveness of the co-selection of forms. This revealed that, although a greeting form and farewell form may independently have relatively low likelihood ratios, such as Arnold's greeting of recipient name followed by a colon and his *Thx*(+Name) farewell (with likelihood ratios of 10 and 9 respectively), their co-occurrence in an email result in that email being far more likely to have been written by Arnold, with a combined likelihood ratio of 90.

This analysis has focused on emails, and the reality is that some genres or text-types have less well-defined genre conventions, which would cause problems for a genre-focused approach such as this. Further, this study has worked with the ideal in terms of a relevant collection of reference data. Although the situation is improving, there remains a lack of population statistics available for authorship analysis (Coulthard 2010:483). Even within this study, although ESEARC goes far in meeting the criteria of offering a 'relevant population' of writers for comparison, there can always be the argument that a reference corpus could be,

and should be, larger in order to provide reliable evidence of the expectancy, rarity and distinctiveness of particular linguistic forms. The drawback of this is that statements of likelihood here are limited to likelihood in relation to a 'relevant' population, rather than the ideal 'general population'. In addition, although this paper has used likelihood ratios as a tool in an empirical investigation into the distinctiveness of particular linguistic choices, Coulthard (2010:483) highlights the potential difficulties in translating their significance to lay juries in forensic casework.

Further research will involve increasing the number of variables considered in the analysis, as well as the explicit and systematic analysis of the ways in which choices of greetings and farewells interact with the sociolinguistic context of each email, such as the relationship between participants, the purpose and topic of the email, and the point in the overall conversation in which the email occurs. This will provide a more full understanding of how particular authors behave linguistically in email, and whether this behaviour is unique or idiolectal.

Overall, the results of this empirical investigation have shown that, in the first instance, stylistic variation and linguistic choices within the conventions of email greetings and farewells are powerful in distinguishing between the writings of a small subset of (in this case) four authors. Moreover, such greeting and farewell forms can be very distinctive of a particular author even when tested against a reference corpus of socially similar authors, and those which are initially low-frequency may be most rare, marked and distinctive. These findings support the suggestion that a genre-focused approach to identifying and analysing idiolect has strong potential for authorship research. As a result, this paper can at least serve as a launching point for more focused empirical research into idiolect, and offers both theoretical and methodological contributions, as well as a baseline of population results, for forensic authorship casework involving emails.

8 References

Abbasi, A. and Chen, H. (2005) Applying authorship analysis to extremist-group web forum messages. *IEEE Intelligent Systems* 20(5): 67–75.

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

Biber, D. (1993) The multi-dimensional approach to linguistic analyses of genre variation: An overview of methodology and findings. *Computers and the Humanities* 26(5/6): 331–345.

Biber, D. and Conrad, S. (2009) *Register, Genre and Style.* Cambridge: Cambridge University Press.

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

Butters, R. (2012) Retiring President's closing address: ethics, best practices, and standards. In S. Tomblin, N. MacLeod, R. Sousa-Silva and M. Coulthard (eds.) *Proceedings of the Tenth International Association of Forensic Linguists' Biennial Conference*, Aston University, Birmingham, 351–361. [online]. From: www.forensiclinguistics.net

Cohen, W.W. (2009) *Enron Email Dataset*. Retrieved October 2010 from: http://www.cs.cmu.edu/~enron/.

Corney, M., Anderson, A., Mohay, G., and de Vel, O. (2001) *Identifying the authors of suspect Email.* Retrieved December 2001 from: http://eprints.qut.edu.au/8021/1/CompSecurityPaper.pdf

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. (2004) Author Identification, Idiolect, and Linguistic Uniqueness. *Applied Linguistics* 24(4): 431–447.
Coulthard, M. (2010) Experts and opinions: In my opinion. In M. Coulthard and A. Johnson (eds.) *The Routledge Handbook of Forensic Linguistics* 473–486. London: Routledge.
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.
Crystal, D. (2001) *Language and the Internet*. Cambridge: Cambridge University Press.
Crystal, D. (2008) *Txtng: The Gr8 Db8*. Oxford: Oxford University Press.
De Beaugrande, R. (1998) Language and Society: The Real and the Ideal in Linguistics, Sociolinguistics, and Corpus Linguistics. Retrieved April 2012 from http://www.beaugrande.com/jsocioling.htm. Also in *Journal of Sociolinguistics* 3(1): 128–139.
de Vel, O., Anderson, A., Corney, M., Mohay, G. (2001) Mining e-mail content for author identification forensics. *Association for Computing Machinery Sigmod Record* 30(4): 55–64.
Gains, J. (1998) 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.
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.
Halliday, M.A.K. and Hasan, R. (1989) *Language, Context and Text: Aspects of Language in a Social-Semiotic Perspective*. Oxford: Oxford University Press.
Holmes, Janet. 2001. *An Introduction to Sociolinguistics* (2nd edn.). Harlow: Longman.
Hymes, D. (1974) *Foundations in Sociolinguistics: An Ethnographic Approach*. London: Tavistock.
Jakobson, R. (1956) *Fundamentals of Language*. Hague: Mouton Press.
Johnson, A. (2012) Applying forensic linguistics in professional settings: Implications for research. Paper presented at the *1st Inter-university PhD Seminar on Forensic Linguistics (University of Leeds & IULA/Universitat Pompeu Fabra*), Universitat Pompeu Fabra, Barcelona. 30 March 2012.
Johnstone. B. (2009) Stance, Style, and the Linguistic Individual. In A. Jaffe (ed.) *Stance: Sociolinguistic Perspectives.* 29–52. Oxford: Oxford University Press.
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. Peter Lang: Frankfurt am Mein.
Kuhl. J. (2003) The Idiolect, Chaos, and Language Custom Far From Equilibrium: Conversations in Morocco. Unpublished PhD Thesis, Athens, Georgia.
Lan. L. (2000) Email: a challenge to Standard English? *English Today* 16(4): 23–29.
Lyons, J. (1968) *Introduction to Theoretical Linguistics*. Cambridge: Cambridge University Press.
MacLeod, N. and Grant Tim. (2012) Whose Tweet? Authorship analysis of micro-blogs and other short form messages. In S. Tomblin, N. MacLeod, R. Sousa-Silva and M. Coulthard (eds.) *Proceedings of the Tenth International Association of Forensic Linguists' Biennial Conference*, Aston University, Birmingham, 210–224. From: www.forensiclinguistics.net

McGee, S. (2002) Simplifying likelihood ratios. *Journal of General Internal Medicine* 17(8): 647–650.
Mollin, S. (2009) "I entirely understand" is a Blairism: The methodology of identifying idiolectal collocations. *International Journal of Corpus Linguistics* 14(3): 367–392.
Scott, M. (2008) *WordSmith Tools version 5*. Liverpool: Lexical Analysis Software.
Smith, D.J., Spencer, S. and Grant, T. (2009) Authorship analysis for counter terrorism Unpublished Research Report, QinetiQ/Aston University.
Solan, L. (2012) Ethics and method in forensic linguistics. In S. Tomblin, N. MacLeod, R. Sousa-Silva and M. Coulthard (eds.) *Proceedings of the Tenth International Association of Forensic Linguists' Biennial Conference,* Aston University, Birmingham, 362–368. From: www.forensiclinguistics.net
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.
Waldvogel, J. (2007) Greetings and closings in workplace email. *Journal of Computer-Mediated Communication* 12(2): 456–477.
Wales, K. (2001) *A Dictionary of Stylistics*. Harlow: Longman.
Wardhaugh, R. (2006) *An Introduction to Sociolinguistics* (5th edn.). Oxford: Blackwell.
Woolls, D. (2012) *Description of CFL extraction routines for CMU Enron Sent email database.* Retrieved March 2012 from: http://www.cflsoftware.com/CFL_CMU_Enron_Sent_email_Extraction.mht