This is a post-review, pre-publication (post-print) version of the paper: Wright, D. (2017) Using word n-grams to identify authors and idiolects: A corpus approach to a forensic linguistic problem. To appear in the *International Journal of Corpus Linguistics* 22(2).
https://benjamins.com/#catalog/journals/ijcl.22.2.03wri/details

Using word n-grams to identify authors and idiolects

A corpus approach to a forensic linguistic problem

David Wright
Nottingham Trent University

Forensic authorship attribution is concerned with identifying the writers of anonymous criminal documents. Over the last twenty years, computer scientists have developed a wide range of statistical procedures using a number of different linguistic features to measure similarity between texts. However, much of this work is not of practical use to forensic linguists who need to explain in reports or in court why a particular method of identifying potential authors works. This paper sets out to address this problem using a corpus linguistic approach and the 176-author 2.5 million-word Enron Email Corpus. Drawing on literature positing the idiolectal nature of collocations, phrases and word sequences, this paper tests the accuracy of word n-grams in identifying the authors of anonymised email samples. Moving beyond the statistical analysis, the usage-based concept of entrenchment is offered as a means by which to account for the recurring and distinctive production of idiolectal word n-grams.

Keywords: forensic linguistics, idiolect, authorship attribution, entrenchment, Enron

1. The linguistic individual, corpora and forensic linguistics

'Idiolect' is a well established concept in linguistics, yet the individual is rarely the focus of linguistic enquiry. There are many possible reasons for this, but perhaps the main deterrent to the study of idiolect is the practical difficulties in doing so. Bloch (1948: 7) coined the term 'idiolect' to refer to "not merely what a speaker says at one time: it is everything that he *could* say in a given language" (original emphasis).

Clearly, the task of collecting anything that a person *could* say is an impossible one. However, recent work in corpus linguistics that has put the individual at the centre of their investigations has narrowed the goal posts set out by Bloch (1948) by analysing the linguistic output that individual speakers or writers actually produce (e.g. Coniam 2004, Mollin 2009, Barlow 2013). These studies use smaller, specialised corpora to systematically examine idiolectal variation that is masked or buried in traditional large-scale reference corpora.

The field which stands to benefit the most from the empirical investigation of idiolect is forensic linguistics, and in particular forensic authorship attribution. Authorship attribution is the process in which linguists set out to identify the author(s) of disputed texts using identifiable features of linguistic style, ranging from word frequencies to preferred syntactic structures. In a forensic context, the disputed texts under analysis are potentially evidential in alleged infringements of the law or threats to security. Such texts can include abusive emails, ransom notes, extortion letters, falsified suicide notes, or text messages sent by a person acting as someone else. In the most straightforward case, the analysis requires the linguist to analyse the style(s) exhibited in the "known" writings of the suspect or candidate authors involved in the case. Attention then turns to the disputed document(s), as the linguist compares the writing style of the text(s) in question and examines the extent to which it is similar or consistent with the known writing style of one (or more) of the suspects. The linguist may then express an opinion as to how likely it is that the disputed text is or is not written by one of the suspects. Such an analysis relies on a theory of idiolect (Coulthard 2004: 431), or at least depends on the *consistency* and *distinctiveness* of the styles of the individuals involved (Grant 2013: 473).

There are a small number of studies and cases in which corpora or corpus methods have been used to attribute forensic texts to their authors. Svartvik (1968) uses a corpus approach to analyse a set of disputed witness statements in a murder case. Coulthard (1994) uses specialised corpora of ordinary witness statements and police statements, along with the much larger spoken element of the COBUILD corpus, in his seminal analysis of the disputed Derek Bentley statement. Coulthard (2004) reports another case in which the internet was used to investigate the author-distinctiveness of twelve lexical items co-selected in one text in the capturing of the Unabomber. Despite

the success of corpus approaches in these cases, few have pursued the utility of corpus linguistics in forensic research. Kredens (2002) is the earliest exception, using a corpus-approach to comparing the idiolects of two English musicians, Robert Smith (*The Cure*) and Steven Morrisey (*The Smiths*). Larner (2014) is an exception too, with his work on identifying idiolectal preferences for formulaic sequences in personal narratives, while Grant (2013) uses a corpus method to identify lexical variation in text messages central to a murder investigation, and Wright (2013) and Johnson and Wright (2014) employ corpus techniques in the analyses of author-distinctive language use in a corpus of business emails. This study continues to develop the use of corpus methodologies in the investigation of idiolect and the attribution of disputed texts in a forensic context. There are two parts to the analysis in this paper. The first part reports the results of an authorship attribution experiment using 'word n-grams' as style markers. The second part focuses on one author as a case study and examines the n-grams which were most useful in identifying his disputed texts, discussing their nature and their implications for the theory of idiolect and forensic authorship analysis.

2. Word strings as features in authorship analysis

Most of the work in authorship attribution is from computer science and computational linguistics. The last two decades have seen an explosion in the number of different linguistic features that have been used to discriminate between authors and attribute samples of writing to their correct author. These range from average word/sentence length, vocabulary richness measures and function word frequencies, to word, character and part-of-speech sequences (Stamatatos 2009**)**. This research is unquestionably valuable; there is now little doubt that by using a combination of linguistic features and a sophisticated machine learning technique or algorithm we are able to successfully identify the most likely author of a text. What we cannot do with the same confidence, however, is explain *why* these methods work. As Argamon & Koppel (2013: 299) comment, "in almost no case is there strong theoretical motivation behind the input feature sets, such that the features have clear interpretations in stylistic terms". Herein lies the problem for forensic linguists, who must be able to say why the features they

describe might distinguish between authors (Grant 2008: 226). We cannot expect lay decision makers such as judges and jurors to understand methods and results which we cannot explain ourselves.

Word strings offer one possible remedy. Sinclair's (1991: 109) 'idiom principle' holds that a language user "has available to him or her a large number of semi-preconstructed phrases that constitute single choices". In the twenty-five years since the idiom principle was first introduced, there has been considerable research attention paid to word strings, with different studies naming, identifying and characterising them in different ways depending on the research goals at hand (Biber et al. 2004: 372; Wray 2002: 9). Despite using different terminology, originating from different theoretical positions, and developing from different disciplines of linguistics, it is possible to identify a common feature in previous work on word strings: their individual nature. The following sections give an overview of some of the prominent theories regarding the individuality of word strings, their relationship with routine communicative events, and the existing empirical evidence of their individual nature. Finally, focus shifts to how the present study builds upon this previous work by utilising word n-grams as a means of attributing disputed texts and identifying idiolectal variation.

2.1. Word strings, routine and the individual

Hoey (2005: 8) argues that "we can only account for collocation if we assume that every word is mentally primed for collocational use". Hoey (2005: 15) draws on Firth's (1957) notion of 'personal collocations', emphasising that "an inherent quality of lexical priming is that it is personal" and that "words are never primed *per se*; they are only primed for someone". He argues that everyone's primings are different and that everyone's language is unique as a result of different linguistic encounters, different parents, friends and colleagues (Hoey 2005: 181). This is a premise shared by Barlow (2013: 444), as he points out that from a usage-based perspective, an individual's cognitive representation of language is influenced by "the frequency of the different expressions and constructions encountered by the speaker." This idea that differing socio-historical linguistic backgrounds lead to differences in repertoires of choice

appears to be acceptable to forensic linguists as a means by which to account for inter-author variation (Nini & Grant 2013: 175).

Wray (2002: 9) introduces 'formulaic sequences' as sequences of words (or other elements) which appear to be pre-fabricated and retrieved whole from memory at the time of use. The term was coined as a coverall, to consolidate "any kind of linguistic unit that has been considered formulaic in any research field" (Wray 2002: 9). Although Wray (2008: 67) marks a clear distinction between formulaic sequences and lexical priming insofar as what constitutes the "fundamental currency of processing", she too emphasises individual variation. While particular sequences are formulaic "in the language" and are shared across the speech community, she argues that "what is formulaic for one person need not be formulaic for another" (Wray 2008: 11). Schmitt et al. (2004) argue something similar. They ran oral-response dictation tasks to test whether corpus-derived recurrent word clusters are stored holistically as psychologically "real" formulaic sequences for native and non-native speakers of English. Results varied, with native speakers performing better than non-natives. While the authors emphasise that the dictation task is an indirect measure of holistic storage (Schmitt et al. 2004: 147), they did report that some recurrent clusters are "highly likely" to be formulaic sequences (such as *go away* and *I don't know what to do*), while others are "quite unlikely" to be (such as *in the same way as* and *aim of this study*) (Schmitt et al. 2004: 138). Between these, they state, are clusters that will be formulaic for some people and not others; "it is idiosyncratic to the individual speaker whether they have stored these clusters or not" (Schmitt et al. 2004: 138). Furthermore, they offer an argument that echoes Hoey's (2005: 181) and Barlow's (2013: 444) explanations for idiolectal collocational preferences. They propose that as part of their idiolect, "it is reasonable to assume that individuals have their own unique store of formulaic sequences based on their own experience and language exposure" (Schmitt et al. 2004: 138).

There exists a relationship between such recurring word sequences and the specific communicative purposes they fulfil. Some argue that this relationship is pervasive through language, such that "we start with the information we wish to convey" in a given situation, and then we "haul out of our phrasal lexicon some patterns that can provide the major elements of this expression" (Becker 1975: 62). Others (e.g.

Kuiper 2004: 41, 45) have argued that in conventionalised contexts particular 'formulae' are "keyed to particular contexts and roles within those contexts". Before Wray's (2002) introduction of 'formulaic sequences', Nattinger & DeCarrico (1992: 1) coin the term 'lexical phrases' as being "chunks of language of varying length" which "occur more frequently and have more idiomatically determined meaning than language that is put together each time". Integral to the concept of lexical phrases is their functional role. Nattinger & DeCarrico (1992: 36) state that the use of lexical phrases is governed by "principles of pragmatic competence", which "select and assign particular functions to lexical phrase units". However, individual variation remains crucial. Like Wray (2002), Nattinger & DeCarrico (1992: 39-40) argue that while many of these are general phrases used by almost everyone in the speech community, such as *how do you do* and *how are you*, some may be "idiosyncratic phrases that an individual has found to be an efficient and pleasing way of getting an idea across." Wray's (2002) view of formulaic language also aligns with this situationally-influenced use of word sequences, as she argues that formulaic language is a dynamic response "to the demands of language use, and, as such, will manifest differently as those demands vary from moment to moment and speaker to speaker" (Wray 2002: 5). This relationship between routine language use and individuality is central to some usage-based theories of grammar. Langacker (1988: 59) states that with repeated use, a once novel lexico-grammatical structure "becomes progressively entrenched, to the point of becoming a unit" and that "through repetition, even a highly complex event can coalesce into a well-rehearsed routine that is easily elicited and reliably executed" (Langacker 2000: 3). Schmid (2016) presents a detailed discussion of the concept of 'entrenchment' and identifies a range of factors which determine the entrenchment of particular sequences, including word strings. He claims that while frequency of occurrence influences the entrenchment process, frequency is simply an "approximation of repeated use and exposure by individual speakers taking place in concrete situations", and that "it is only in communicative situations that replication and subsequent propagation" can take place (Schmid 2016: 18-9). He highlights that entrenchment relates to the minds of individual and therefore is "more or less by definition subject to individual, speaker-related differences" (Schmid 2016: 21). He goes on to explain that the sources of these differences are "hidden in the exposure and usage histories of individual speakers",

which are influenced by social variables including region, gender, education and training, as well as by "personal routines and experiences" (Schmid 2016: 21).

There is some agreement across different disciplines in linguistics that particular word strings are functionally tied to specific recurring communicative contexts, routines and purposes. While some of these routines and resultant word strings are shared across the speech community, others may be more personal, or even unique, to individuals. Therefore, word strings offer the authorship analyst a linguistic feature for which there is some theoretical consensus that can help explain differences between authors. To date, however, there is only a small body of empirical evidence supporting the idiolectal nature of word strings and, by extension, their applicability in forensic authorship analysis.

2.2. Empirical evidence for idiolectal word strings

Most of the research investigating idiolectal patterns of word strings has been produced by corpus linguists. Mollin (2009) analyses a 3 million word corpus of speech and writing of the former Prime Minister of the UK, Tony Blair, focusing on his distinctive use of maximiser collocations such as *entirely reasonable* (maximiser + adjective), *extremely closely* (maximiser + adverb) and *totally accept* (maximiser + verb). Her aim is to identify those collocations that were "truly typical of the individual" (Mollin 2009: 367). Comparing the Tony Blair data with the British National Corpus she identifies 42 maximiser collocations that were over-proportionately used by Tony Blair. After measuring his preference for these forms over synonymous alternatives (e.g. *absolutely central* vs. *fully central*), and eliminating those which were over-represented in particular registers in the BNC (speech, newspaper style, parliamentary style), she finds 25 maximiser collocations that can be considered truly "typical" of Tony Blair, including *entirely understand, absolutely committed* and *perfectly prepared*. Barlow (2013) compares the use of two and three word strings in the speeches of six White House press secretaries, using a corpus of approximately 3.6 million words. After presenting the differences in frequency with which the six speakers use the most common word bigrams in the corpus (e.g *and the, the president, I think),* Barlow (2013:

455) goes on to show how samples of 200,000 words from each of the press secretaries cluster together when bigrams are used as the basis for a correspondence analysis. This evidence, he argues, along with comparisons of trigram use (e.g. *move forward on* and *in terms of*) and part-of-speech bigrams, shows that there is an inbuilt "preference for familiar routines leading to a consistency in frequency of usage of language expressions by individual speakers" (Barlow 2013: 472). In a forensic context, Coulthard (2004: 441) argues that the longer a sequence of words is then the less likely it is that any two writers will use that identical sequence in two separate texts. He demonstrates this by testing the uniqueness of the strings *I picked something up like an* and *I asked her if I could carry her bags*. Coulthard (2004) enters these strings into Google, starting from two words and adding an additional word each time. By the time the strings became six to eight words long, the search returned zero results, and Coulthard (2004: 42) argues that "rarity scores like these begin to look like the probability scores DNA experts proudly present in court."

Despite the methodological differences across these studies, Coulthard (2004) Mollin (2009) and Barlow (2013) all provide corpus-derived evidence that supports of the notion of idiolectal word strings. In an authorship context, however, where word strings have been used to attribute texts to their correct authors, they have returned mixed results. Hoover (2002) uses cluster analysis to determine whether literary texts by the same author could be distinguished from those by different authors using the most frequent two-word sequences occurring across the corpora (this ranged from the 50 to the 800 most frequent). Ultimately, Hoover (2002: 176) finds that frequent word sequences are more accurate in clustering texts by the same author than the most frequent single words, which have typically been considered one of the most effective features for authorship analysis. Similarly, Coyotl-Morales et al. (2006) use "maximally frequent word sequences" of between one and three words in length to attribute samples of poetry to their correct authors, and using classification algorithms, they report an accuracy rate of 83%. As in Hoover's (2002) study, this performance was better than function words, which they argue "do not help capture the writing style from short documents" (Coyotl-Morales et al. 2006: 7). In a forensic context, Juola (2013) takes a slightly different approach. Using *all* of the three-word sequences which appear in his data, rather than only the most frequent, he attributes a set of ten anonymously-written

anti-government articles to the person who claimed authorship of them in a deportation case. In order to demonstrate that these disputed documents were written by the person in question, Juola (2013) compares these ten disputed documents with a set of ten articles known to have been written by the author and five additional sets of articles totalling 160 texts written by different named authors in the same language. On the basis of three-word sequences, the disputed documents were measured as being more similar to the author's known articles than to any of the other five distractor authors, and this provides evidence to support the author's claim that they had written the disputed articles.

In other studies word n-grams have not fared so well. Grieve (2007: 263) evaluates the success of collocations in the attribution of newspaper columns to their correct author and finds that they performed poorly. Two-word and three-word collocations achieved a success rate of 75% and 53% respectively when distinguishing between authors. In fact, the three-word collocations were the least successful of the many features tested in his study. In comparison, character-level n-grams performed far better, with two, three and four character strings distinguishing between two authors with accuracies of 93% and 94%. This finding aligns with that of Sanderson & Guenter (2006: 9) who also find that character sequences generally outperform word sequences in their attribution of newspaper texts written by 50 journalists. Something that these authorship studies have in common is that the readers are not shown any of the specific word strings that were useful in the attributions. This contrasts with Coulthard (2004), Mollin (2009) and Barlow (2013) where the idiolectal nature of a precise set of collocations is tested. Also, there is often little or no explanation offered as to why word sequences were or were not useful in these studies. An exception to this in a forensic authorship attribution context is Larner (2014), who tests the usefulness of formulaic sequences (in Wray's [2002] terms) as markers of authorship. Larner (2014: 10) constructs a list of 13,412 "clichés", "idioms", "proverbs", "similes" and "everyday expressions and sayings" defined as such in various online sources. Of these 13,412 "formulaic sequences", 301 were found in the 100 personal narratives he had collected from twenty different authors, including phrases such as *in the end*, *at least*, *go back* and *in fact*. Using Jaccard's co-efficient to measure similarity between texts, Larner (2014: 13) finds that in his corpus "texts produced by the same author are more similar

in their use of formulaic sequence types than text by different authors." However, in terms of using these formulaic sequences to identify the author of a disputed text, he concludes that "neither the type of formulaic sequences nor the overall count of formulaic words enables the attribution of a text to its author" (Larner 2014: 18). Nevertheless, Larner (2014) presents the first move to explicitly investigate formulaic sequences, in the strictest sense, as a marker of authorship. In a way, this represents almost the antithesis of other authorship studies that have utilised word strings. Whereas previous work has produced good attribution results using word strings but offered no theoretical explanation for those results, Larner (2014) adopts a strongly theoretically informed feature set but produces more conservative attribution results. The present study aims to combine these two approaches by first pinpointing the word strings useful in attributing texts to their correct authors, and subsequently presenting a theoretical argument as to why they are useful.

2.3. 'Word n-grams' in this study

Given that there is some theoretical explanation as to why individuals vary in their use of word strings and some, albeit limited, evidence of such idiolectal variation from corpus linguistics and authorship studies, word strings are an ideal candidate for use by forensic linguists. This study, therefore, aims to harness and test their potential. In order to do this, the method used here captures all word strings, between 2 and 6 words in length, in known and disputed sets of texts. In this study, 'word n-grams' is the term used to refer any string of *n* words in length, with no *a priori* assumptions being made regarding their frequency or holistic storage. 'Word n-gram' is an operational term used to refer to strings of words (Juola 2008: 265) which, to borrow Wray's (2002: 9) term, do not carry any theoretical "baggage". The argument here is not that word n-grams hold status as a "special kind" of word sequence akin, for example, to 'formulaic sequences' (Wray 2002) or 'lexical phrases' (Nattinger & De Carrico 1992), but that they offer an objective way of capturing linguistic output of individuals and measuring similarity between texts. Once identified, the word n-grams most useful in attributing

set of texts to their authors can be interpreted in light of the existing theory discussed here.

3. Methodology

This paper comprises two analysis sections. The first (Section 4) reports the results of an attribution experiment which tests the effectiveness of word n-grams in identifying authors. In the experiments, random samples of authors' emails were extracted from their set and anonymised. These 'disputed' samples were then compared against the email sets of the candidate authors on the basis of the number of n-grams they share, to observe which author the method identifies as being responsible for writing the disputed sample. The second part of the analysis (Section 5) focuses on one author—Gerald Nemec—and examines the word n-grams that were useful in correctly identifying him as the author of his samples. This second section goes beyond the statistical results of the attribution experiments, and explores precisely which word n-grams are most distinctive of Nemec's style, and how they offer an insight into his idiolectal preferences. The following Sections 3.1 and 3.2 detail the corpus used for the analyses and the procedure of the attribution experiment.

3.1. The Enron Email Corpus

Enron is a former American energy company which filed for bankruptcy in late 2001 following a now infamous accounting scandal. In 2003, a database of 1.6 million Enron documents, including employees' emails, was released into the public domain by the Federal Energy Regulatory Commission. After Enron employees requested around 140,000 documents be redacted, the final database contained around half a million messages sent and received by Enron employees. The vast majority of these employees were not involved in any criminal activity, and the purposes of this study are not to investigate the procedural behaviour of any individuals; the interest here is entirely linguistic. Various versions of the data are available online, but the one drawn upon

here is that collected and prepared by Carnegie Mellon University (CMU) (Cohen 2009). For the present study, the CMU set has been cleaned and optimised for authorship analysis, removing any duplicate emails, email threads and irrelevant metadata. For the purposes of the present study, only emails sent (rather than received) by Enron employees are included. Each email in the set looks like that in Figure 1. In its totality, the corpus used in this study comprises 176 authors, 63,369 emails and 2,462,151 tokens.

```
<Date: Mon, 1 Oct 2001 15:04:03 -0700 (PDT)>
<From: monika.causholli@enron.com>
<To: cecil.stapley@enron.com>
<Subject: Pulp facts>

Cecil,
can you please take a look at this presentation
and see if numbers make sense?
thanks,
Monika
```

Figure 1. Sample email from the Enron Email Corpus

The corpus is especially suited to authorship analysis for a number of reasons. Firstly, it is naturally-occurring data, rather than being elicited especially for authorship purposes. We can be sure with some degree of certainty that the person's account from which each email is sent is the 'sole executive author' (Love 2002: 43) of the text. The emails are not likely to have been subject to any editorial intervention, for example, which may compromise the style exhibited in the text. One can identify the authors in the Enron Email Corpus as representing a 'community of practice' (Eckert & McConnell-Ginet 1998: 490); all of the authors work for the same company, they are writing using the same medium and in the same text-type, and they are all writing at the same time. Working with emails is also beneficial because forensic cases increasingly involve digital texts (such as emails) containing threatening, abusive, or defamatory material (Coulthard et al. 2011: 538). Finally, in terms of a dataset to analyse idiolect, the Enron corpus offers a contrast to the corpora used in Mollin (2009) and Barlow (2013) as it comprises written rather than spoken data.

3.2. The authorship attribution experiment

In the attribution experiment, random samples of authors' emails were extracted from their set and anonymised, and then the method attempted to correctly identify the author of those samples. Twelve authors were chosen from whom the samples were taken. These twelve authors were selected on the basis of a number of criteria. First, they are all men. Some studies in author profiling have found word sequences to be useful in predicting the biological sex of the writer (e.g. Mikros 2012). This study, however, is not concerned with any potential sex-related variation, and so the sex of the authors was kept constant. Second, between them they have three different roles within the company (four traders, four lawyers, four managers). Third, they have a range of different dataset sizes, from an author with a sub-corpus of 91,621 tokens (2,295 emails) to an author with only 6,042 tokens (467 emails). At first, it may seem counter-intuitive in a study of idiolect to include authors who have different jobs. This will lead to topic and register differences across the corpus and therefore necessarily produce difference in linguistic output across the authors. However, as will be clear from the analysis and discussion below, attempting to disentangle the identities of a person (including their job) from any discussion or analysis of their idiolect, is impossible. Another justification, although slightly more expedient, is that in a forensic case the analyst is not given the luxury of a balanced, representative and controlled corpus. Rather, what they receive from the police or solicitors is often "any old collection of texts" (Cotterill 2010: 578). Finally, the texts involved in forensic cases are often unhelpfully short, so there is a need in the field to test methods on small samples. Having authors with a range of dataset sizes facilitates this challenge. For each of the twelve authors, ten random samples of 20%, 15%, 10%, 5% and 2% of their emails were extracted and anonymised, giving a total of 50 samples per author. These samples are discrete and each contains different randomly selected e-mails, rather than the 15% sample being a reduced version of the 20% sample, for instance. Table 1 shows the number of emails and mean number of tokens that make up each sample size for each of the twelve authors. Taking proportions of authors' email sets is preferable to taking a fixed number of emails for each author in this case given the differences in dataset sizes of the twelve authors. For example, if 100 emails were used as a sample across all authors, this would account for only 4.4% of

Table 1. Size of samples in experiment*

	Total		20%		15%		10%		5%		2%	
Author	**Emails**	**Tokens**	**E**	**MT**	**E**	**MT**	**E**	**MT**	**E**	**MT**	**E**	**MT**
Germany	2,295	91,621	459	13,436	344	10,116	229	6,820	114	3,367	45	1,317
Kaminski	2,297	54,498	459	9,362	344	6,909	229	4,656	114	2,200	45	955
Nemec	1,466	58,911	293	9,224	219	7,047	146	4,963	73	2,327	29	998
Steffes	1,202	35,668	240	6,642	180	5,126	120	3,033	60	1,400	24	655
Lavorato	1,112	25,320	222	4,600	166	3,360	111	2,217	55	1,031	22	450
Arnold	1,036	26,659	207	4,633	155	3,281	103	2,481	51	1,272	20	449
Haedicke	800	19,622	160	3,294	120	2,309	80	1,591	40	780	16	286
Farmer	772	24,502	154	4,208	115	3,213	77	2,016	38	1,030	15	350
Dorland	502	14,605	100	2,464	75	1,935	50	1,279	25	655	10	280
Derrick	467	6,042	93	951	70	712	46	485	23	253	9	89
Allen	357	16,710	71	2,575	53	2,001	35	1,267	17	609	7	255
Zipper	242	7,912	48	1,281	36	1,090	24	554	12	259	4	151

* (E = emails, MT = mean tokens per sample)

Germany's emails, but 41.7% of Zipper's. Using percentages, therefore, counteracts these size discrepancies and makes the samples, and therefore results, more comparable across authors.

Overall, there were 600 samples (12 authors, five sample sizes, ten samples each) which represent the 'disputed' texts for which the author is to be identified. Despite containing many emails, each of the 600 samples is essentially treated as one text. Creating samples of an author's writing by combining a number of texts they have written, rather than individual texts, is a common practice in authorship studies (e.g. Luyckx & Daelemans 2011, Juola 2013, Stamatatos 2013). In order to identify the most likely author of these samples, they were measured for similarity against the remaining emails of the author from which they had been extracted, and the full email sets of the other 175 authors. These sets against which the disputed samples are compared represent the 'known' writings of the authors in the corpus. As with the disputed samples, these sets of known emails were each treated as one text. Having 176 'candidate authors' for a disputed sample is unusual in authorship research, which normally tests methods on far fewer authors, and sometimes only two (Luyckx & Daelemans 2011: 37). The comparisons of the disputed samples and the known sets were pairwise, so at any one time the disputed sample was being compared with one known set (Figure 2). The similarity between the disputed samples and known sets in

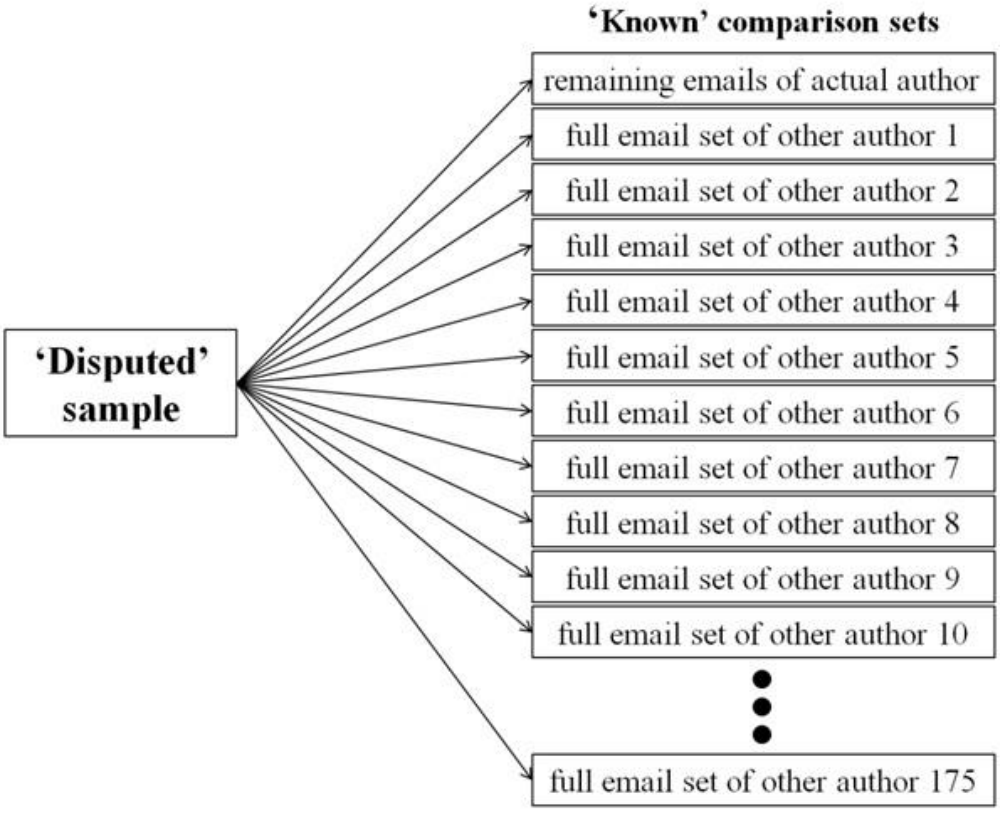

Figure 2. The pairwise comparison process measuring the disputed sample against 176 sets of known writing

this experiment is measured using Jaccard's similarity co-efficient, a statistic which has its origins in ecology but has recently been adopted by forensic linguists (Grant 2013, Juola 2013, Johnson & Wright 2014, Larner 2014). Jaccard is a simple calculation which takes into account the number of different features shared between two datasets (A∩B) as a proportion of all the features in the two sets combined (A∪B) using the formula:

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

Jaccard is a binary correlation analysis in that it hinges on the appearance or non-appearance of a particular feature in the two samples compared rather than how frequently it occurs. In this experiment, each word n-gram of a given length (i.e. a 'type') constitutes a 'feature' in its own right, and the disputed and known samples are

compared on the basis of the total number of different n-grams they share, rather than on any individual n-gram on its own. Five different lengths of n-gram are tested separately: bigrams through sixgrams (within sentence boundaries). As an example, consider there were 395 different bigrams that were found in both the disputed sample of an author and their remaining known email set (e.g. *of the*, *you can*, *it is*). Together, the disputed and known sets being compared have a total of 8,767 different bigrams, including those shared. This would mean 395/8,767 would give a Jaccard score of 0.045. Jaccard produces a result between 0 and 1, with 0 indicating that there is no similarity at all between the sets being compared, and 1 meaning that they are identical. In this experiment, the author whose known set returns the highest Jaccard score is measured as being most similar to the disputed sample, and therefore the most likely author of that sample. Therefore, an attribution is considered successful if the highest Jaccard score is returned when the disputed sample and known set being compared belong to the *same* author. If the disputed sample is scored as being most similar to a known set of a *different* author, then this is a misattribution. This procedure follows that successfully employed by Juola (2013), who ranks Jaccard scores for six authors when comparing their writing with a set of questioned documents. It should be noted, however, that other authorship studies investigating smaller datasets and fewer authors (Grant 2013, Larner 2014) have used additional tests to determine the statistical significance of differences in mean Jaccard scores across texts and authors. Such studies have tended to use non-parametric Mann-Whitney U to test whether the mean Jaccard scores for samples with the same author are significantly different from samples with different authors. However, these studies have tested with far fewer authors (sometimes only two), so that the mean Jaccard scores are more comparable across samples. Given that in any one test in this experiment there is one "correct" author and 175 "incorrect" authors the means are less comparable, and the power of the statistic would be too diminished to be of any use. Therefore, such significance tests are not used here.

The entire process described here of randomly sampling and extracting authors' emails, the comparison of disputed and known sets, and the Jaccard calculations are performed using a bespoke piece of Java-based linguistic analysis software developed by Woolls (2013) called *Jaccard N-Gram Lexical Evaluator* (*Jangle*). As well as running the similarity tests, the program allows the analyst to identify precisely which

n-grams were shared between the known and disputed data and accounted for the successful attribution of the disputed sample. It is this qualitative investigation which is addressed in the second part of the analysis, which also includes concordance analyses performed using *Wordsmith Tools* (Scott 2008).

4. Attribution results

In the attribution experiment, there were a total of 3,000 tests in which the author of a disputed sample was to be identified (12 authors, ten samples of five different sizes, five n-gram lengths). Across all of these tests, the method was successful in identifying the correct author of the sample in 1,913 instances, giving an overall success rate of 64%. This means that in 64% of cases, the highest Jaccard score of similarity was returned when the disputed sample was compared against the known writings of the same author, rather than one of the other 175 candidates. While this success rate is an over-generalisation for a number of reasons (discussed below), it gives some indication that word n-grams do have the potential to be useful markers of writing style and a means by which to measure similarity between known and disputed texts. Stylometric studies (e.g. Grieve 2007, Koppel et al. 2011) consider 70%–75% accuracy in an attribution task to be "successful", "satisfactory", or "passable". The 64% success rate in this experiment is approaching acceptable levels of reliability in stylometric terms, though is not reliable enough for forensic casework. However, there are three points which need to be addressed with regards to the performance of this method: (i) the effect of sample size on accuracy, (ii) the performance of the different n-gram lengths, and (iii) difference in performance across authors.

4.1. Effect of sample size

This method performs better with larger sample sizes than it does with smaller ones (Figure 3). With the smallest samples, no n-gram length achieves higher than 40% accuracy (Table 2). However, by the time the samples are 15% and 20% in size, success

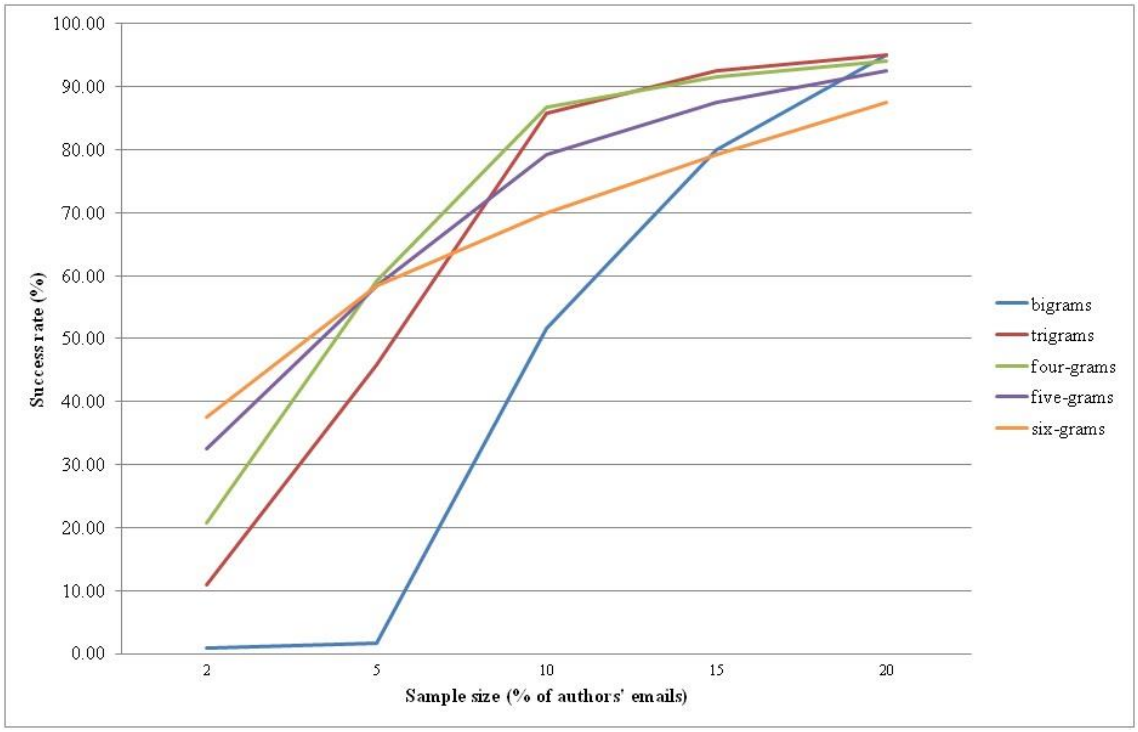

Figure 3. Attribution success rates of all n-gram lengths across all sample sizes

Table 2. Overall success rates across different sample sizes (600 tests per sample size)

samples (tokens)	**bigrams** (n, %)	**trigrams** (n, %)	**four-grams** (n, %)	**five-grams** (n, %)	**six-grams** (n, %)	**total** (n, %)
2% (89-1,317)	1 (0.83)	13 (10.83)	25 (20.83)	39 (32.50)	45 (37.50)	**123 (20.50)**
5% (253-3,367)	2 (1.67)	55 (45.83)	71 (59.17)	70 (58.33)	70 (58.33)	**268 (44.67)**
10% (485-6,820)	62 (51.67)	103 (85.83)	104 (86.67)	95 (79.17)	84 (70.00)	**448 (74.67)**
15% (712-10,116)	96 (80.00)	111 (92.50)	110 (91.67)	105 (87.50)	95 (79.17)	**517 (86.17)**
20% (951-13,436)	114 (95.00)	114 (95.00)	113 (94.17)	111 (92.50)	105 (87.50)	**557 (92.83)**
Total	**275 (45.83)**	**396 (66.00)**	**423 (70.50)**	**420 (70.00%)**	**399 (66.50)**	**1913 (64.00)**

rates exceed 85% and 90% accuracy, which are closer to the reliably rates likely to be suitable in forensic casework. It is not surprising that accuracy decreases with less data. Other authorship studies that have systematically reduced the amount of data used in authorship experiments have also reported drops in success rates (e.g. Grant 2007,

Luyckx & Daelemans 2011, Eder 2015). However, although results are generally poor for 2% samples, the method tested here does attribute some very small individual samples to their correct author. A case in point is the results for Allen. Allen has the third smallest sample sizes in the experiment and his 2% samples range between 108 and 368 tokens. In this experiment, four of these samples are attributed to him using trigrams, five using four-grams, seven using five-grams, and six-grams attribute all but one of these ten very small samples to Allen. Similarly, Derrick has the smallest 2% samples ranging in size between 55 and 145 tokens, yet trigrams and six-grams identify him as the author of one of these ten samples, while four-grams and five-grams attribute three of them to him. The smallest of Derrick's samples to be attributed to him, and the smallest samples to be attributed overall in the experiment, are only 109, 84 and 77 tokens in size (see Johnson & Wright 2014 for a case study on Derrick's style). Therefore, although accuracy rates overall decrease with sample size, this method is still successful in attributing exceptionally small segments of data to their correct authors, even when there are 176 candidate authors in each attribution task.

4.2. Performance of different n-gram lengths

Some lengths of n-gram performed better than others in the experiment (Figure 3 and Table 2). The first thing to note is that the longer n-grams of five and six words in length outperform the shorter ones when attributing smaller samples, but the reverse is the case for larger samples. Each n-gram length was tested on 600 disputed samples across the twelve authors and five sample sizes, and the n-gram length which performed best in the experiment is four-grams, successfully attributing 423 (70.50%) of these 600 samples to their correct author, with success rates as high as 94% with 20% samples. These are followed very closely by five-grams, with a success rate of 420 (70%) out of 600 (Table 2). Authorship and plagiarism studies generally hold that the longer a lexical sequence is, the more likely it is to be idiolectal for the speaker or writer using it. Coulthard (2004: 441-2) argues that, with longer sequences, it is less likely that "the occurrence of this identical sequence in two different texts is a consequence of two speakers/writers coincidentally selecting the same chunk(s) by chance". Therefore, we

might expect that, because they are the longest, six-grams would be the most successful in this experiment. However, while longer measures such as six-grams may be more distinctive, authors may be less likely to repeat them. Four-grams, in contrast, appear to be long enough to be distinctive of individual authors, but also repeated frequently enough by an author that they can be found in both disputed samples and their remaining known emails.

4.3. Performance across authors

Although the method works best with larger sample sizes, and four-grams are the best performing n-gram length overall, the method works better for some authors than others. Each author underwent 250 attribution tests: ten samples of five different sizes were tested using five n-gram lengths. The success of the method in identifying the author of disputed samples ranges from as high as 87.20% accuracy with Lavorato to as low as 8.40% with Zipper (Table 3).

Table 3. Attribution success rates for individual authors

Author	Successful tests (n=250)	Sample size rank
Lavorato	218 (87.20)	5
Allen	210 (84.00)	11
Kaminski	209 (83.60)	2
Arnold	192 (76.80)	6
Germany	183 (73.20)	1
Farmer	173 (69.20)	8
Nemec	172 (68.80)	3
Derrick	167 (66.80)	10
Steffes	135 (54.00)	4
Dorland	117 (46.80)	9
Haedicke	117 (46.80)	7
Zipper	21 (8.40)	12

Furthermore, although the method generally performs better when attributing larger samples (as shown in Figure 3), those authors with the largest sample sizes are not the easiest to identify. The rightmost column in Table 3 shows the authors' rank in the size

of their datasets (out of 12). Germany has the most data of the twelve test authors, and so his sample sizes are consistently the largest, and by some distance. However, he does not have the highest success rate, as better results are achieved for Lavorato, Kaminski, Allen and Arnold, all of whom have (much) smaller samples than him. Meanwhile, Lavorato has only the fifth largest sub-corpus, yet the method works best when attributing his samples. Equally notable is that, in spite of having the third smallest samples in terms of tokens, Allen is correctly identified as the author of these samples in 84% of cases. Similarly, Derrick's samples consistently contain the fewest tokens at each size, yet the approach performs relatively well in attributing these to him, better than Steffes, Dorland and Haedicke, all of whom have more data. This suggests that the particular author being tested is more important than the size of the samples being attributed. Further granularity in the examination of the results reveals that, not only are word n-grams better at identifying the styles of some authors than others, but *particular* length of word n-gram perform best for *particular* authors (Table 4).

Table 4. Best performing n-gram length for each author

Rank	Author	N-gram	Success
1	Lavorato	five-grams & six-grams	49/50
2	Allen	six-grams	49/50
3	Kaminski	five-grams & six-grams	47/50
4	Arnold	six-grams	48/50
5	Germany	six-grams	41/50
6	Farmer	four-grams	38/50
7	Nemec	five-grams	39/50
8	Derrick	four-grams	39/50
9	Steffes	four-grams	32/50
10	Dorland	four-grams	29/50
11	Haedicke	trigrams & four-grams	29/50
12	Zipper	trigrams	6/50

Each n-gram length was used in 50 tests for each author (ten different samples of five sizes). Five-grams and six-grams correctly identified Lavorato as the author of his disputed samples in all but one of those 50 tests, making them the joint best performing n-grams for his style. However, five and six-grams were not the best performing lengths

for all the authors, with trigrams and four-grams prevailing for others. Based on these results, it might be argued that some authors' idiolects, such as Haedicke's and Zipper's, are manifest and are identifiable in shorter collocational sequences than for authors such as Lavorato, Allen, Kaminski Arnold and Germany. Ultimately, however, the difference in accuracy rates across all twelve authors indicates that while idiolect is observable through combinations of distinctive word n-grams for some people, this may not be the case for all people. Therefore, although the results of this method are promising, one linguistic feature, or any finite set of linguistic features, is unlikely to be able to reliably identify idiolectal evidence for every author in any given corpus.

This first part of analysis has tested and reported the results of the accuracy of word n-grams in identifying authors of disputed email samples. The second part of the analysis will now move to examine the precise nature of the word strings that have proved useful in identifying one author in particular, Gerald Nemec, and attempt to explain *why* word n-grams work.

5. Identifying idiolectal word n-grams

The word n-grams which account for the higher Jaccard scores of similarity and the successful attribution of samples are those which appear in the disputed email sample and the remaining known emails for that author, while not appearing at all in the known sets of the other 175 candidate authors. That is, they need to be *distinctive* of the author in question. Once such word n-grams can be isolated in a person's data, they represent a pool of word strings idiolectal of that author. Such an analysis will be performed here using Gerald Nemec as a case study. Nemec was senior legal counsel at Enron during the period covered by this corpus (1998–2002). He has been chosen as a case study because his samples were relatively difficult to attribute using this method. Despite having the third largest dataset of the twelve authors tested, only 172 (68.8%) of the 250 tests using his samples saw them successfully attributed, ranking him as the seventh easiest (or fifth hardest) to identify of the twelve authors. The n-gram length examined here is five-grams because it was the best performing measure with his samples overall, and he is the only one of the twelve authors for which five-grams stand alone as being

the most successful measure, suggesting that they capture his style in a way that they do not for the other eleven authors. The samples focused on here will be his 5% samples, ranging between 2,030 and 2,825 tokens, as these were the smallest samples of his which were reliably identified using five-grams; eight out of ten were successfully attributed.

Table 5 displays the five-grams that were found shared between Nemec's disputed 5% samples and the remainder of his emails in three or more of the eight successful tests. In addition, these five-grams are distinctive of Nemec at 'population level' (Grant 2010); they are either not used at all by any of the other 175 authors in the Enron Email Corpus, or they are used by another author, but only once. In other words, they are distinctive five-word strings which characterise Nemec's idiolectal style when compared against his fellow Enron authors. The main pattern that emerges from across these distinctive five-grams is that many of them are related to his job as a lawyer in the company and, in particular, are reflective of his collaborative practice with colleagues of drafting and revising legal documents and agreements. They can be categorised as belonging to three main types: those which refer to a *redlined* version of a contract or document (highlighted in red), those which request a *review* of an attachment (in green), and those which refer to documents such as a *draft*, *term sheet* or *exhibit* (in blue). Detailed attention is going to be focused here on the two most frequently used of these five-word strings: *a clean and redlined version* and *please review and let(')s discuss*, which Nemec uses 24 and 23 times respectively in 1,475 emails. *Redlining* is a term used to refer to the marking up of changes, additions, deletions, approvals and rejections in the drafting of a collaboratively produced document (like "track changes"). Nemec uses the string *a clean and redlined version* to express to his recipient that he is attaching to his email an edited document, as shown in Example (1):

Table 5. Nemec's distinctive five-grams in attributing his 5% samples and their frequency in his data

Five-gram	Freq.	Five-gram	Freq.
a clean and redlined copy	**5**	***is a rough draft of***	**7**
a clean and redlined version	**24**	***is the term sheet with***	**3**
a redline with the changes	**3**	*might want to check with*	3(2)
a rough draft of the	**6**	***move to the new garage***	**2**
and redlined version of the	**4**	*please forward to the appropriate*	4(2)
are clean and redlined versions	**5**	***please prepare the form of***	**2**
as an exhibit to the	**2**	***please review and if acceptable***	**11**
attached are clean and redlined	**6**	***please review and let(')s discuss***	**23**
attached as an exhibit to	**2**	***please review and provide any***	**9**
attached is a clean and	30 (4)	***prepare the form of ca***	**2**
attached is a redline with	**6**	*questions please call me at*	5(2)
attached is a rough draft	**7**	***review and let's discuss***	**11**
attached is the form we	**4**	***the form of ca for***	**2**
attached is the term sheet	**3**	***the term sheet with my***	**3**
be attached as an exhibit	**2**	***to be attached as exhibit***	**5**
clean and redlined version of	**18**	***to make it clear that***	**4**
i am fine with this	5(2)	***to move to the new***	**2**
i am ok with the	5(2)	***with the changes we discussed***	**13**
if you are ok with	4(2)	*you need any further information*	5(3)
is a clean and redlined	**24**	***you please prepare the form***	**2**
is a redline with the	**3**		

(1) <Date: Mon, 13 Nov 2000 07:00:00 -0800 (PST)>
<From: gerald.nemec@enron.com>
<To: michael.legler@enron.com>
<Subject: Michiwest Agreement>

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

There is a good deal of uniformity and consistency with which Nemec uses *a clean and redlined version*. The five-gram is often part of the longer sequence *attached is a clean and redlined version* which appears 18 times (Figure 4).

Figure 4. Concordance display for *a clean and redlined version* in Nemec's emails

Nemec uses *please review and let(')s discuss* (9 times with an apostrophe, 14 times without) to direct his recipient to examine an attachment and invite further communication afterwards, such as in Example (2):

(2) <Date: Wed, 7 Mar 2001 10:11:00 -0800 (PST)>
<From: gerald.nemec@enron.com>
<To: staci.holtzman@enron.com>
<Subject: Revised CSA>
<Cc: mark.knippa@enron.com, chris.hilgert@enron.com>

Staci, Attached is the CSA with my comments redlined. *Please review and let's discuss.*

Again, there is remarkable consistency in the way in which Nemec uses this string. In 17 of the 23 instances, *please review and let(')s discuss* constitutes the whole sentence, and in 15 of these, the sentence appears at the very end of the email (<Message-ID:> indicates the start of a new message) (Figure 5).

N	Concordance
1	place would really make business going forward much more efficient. Please review and let's discuss and your earliest convenience. I can be
2	Gerald (G$) <Message-ID: > Attached is a draft for your review. Please review and lets discuss any comments. After re-reading the
3	an informal requests for documents that ECS received from WFEC. Please review and lets discuss asap. <Message-ID: > Staci, Attached
4	you check with Veronica? <Message-ID: > Mark, Attached is a draft. Please review and let's discuss before you send to Huber.
5	> Attached are the Cascade Documents redlined with my changes. Please review and let's discuss further before forward by to Cascade.
6	> Mark, Attached is a draft of the CMS Letter Agreement. Please review and lets discuss. I am not sure I have the correct CMS
7	prepare a Work Offer under the Moutain West MSA with the attached. Please review and lets discuss if you have questions. <Message-ID: >
8	confirm, does not reflect the revisions needed to the existing master. Please review and let's discuss. <Message-ID: > Debra, Darren Should
9	> Staci, Attached is the CSA with my comments redlined. Please review and let's discuss. <Message-ID: > I would include
10	> Attached is a rough draft of the agreement we discussed last week. Please review and let's discuss. <Message-ID: > How is it going?
11	I will prepare the executables and get them to Eva. <Message-ID: > Please review and let's discuss. <Message-ID: > <To: houston > Last
12	transfer, so slight mods would be required for a payment by check. Please review and lets discuss. <Message-ID: > Ken, Could you
13	I am trying to find local counsel in Pa to review this section. Please review and lets discuss. <Message-ID: > Mark, Please review
14	<Message-ID: > Lisa, I made significant revisions to this form. Please review and lets discuss. <Message-ID: > Barry, Attached are
15	to discuss how this limits us and what flexibility this section needs. Please review and lets discuss. <Message-ID: > Barry, I added a some
16	the restated Huber confirms with the netting language we discussed. Please review and lets discuss. <Message-ID: > Kay, Please see the
17	We can discuss though, if that is really the way you want to go. Please review and lets discuss. <Message-ID: > A note from Bieraugel.
18	document with the insurance mark-ups that I noted on my voice mail. Please review and lets discuss. <Message-ID: > Extension of Kennedy
19	the content of the memo. Craig is an experienced litigator in Denver. Please review and let's discuss. <Message-ID: > Sorry for the delay.
20	3.1(d) - Title to Membership Interest - Rob Taylor <Message-ID: > Please review and let's discuss. <Message-ID: > Jessica, Attached is
21	by Mariner that we discussed. I put a 3 day threshold in the language. Please review and lets discuss. <Message-ID: > I have no idea what
22	to discuss how this limits us and what flexibility this section needs. Please review and lets discuss. <Message-ID: > Let me address your
23	since we are selling a portion of this project off to investors. Please review and lets discuss. Timing unfortunately is not long as

Figure 5. Concordance display for *please review and lets discuss* in Nemec's emails

Both of these repeated word strings are associated with recurring communicative contexts in which Nemec finds himself while performing his occupational role of lawyer. The concordance results in Figures 4 and 5 provide evidence to suggest that Nemec has developed at least one effective and efficient way in which to fulfil the functions of drawing his recipients' attention to a marked-up document, or requesting that they review a document and reply to him. These are not the *only* lexical strings he produces to do these things, as indicated by other similar strings in Table 6 (such as *attached is a redline with*, and *please review and provide any comments*/*input*). However, focus here will remain on the two most frequently employed phrases, for which routine and entrenchment offer tools for interpretation. In Nemec's work, he is routinely faced with communicative situations in which he has to draw attention to a marked-up document, or request that his recipient review a document and reply to him. In the first instance, Nemec chose, presumably, what he considered to be the most appropriate or suitable thing to say to achieve his pragmatic aims all (contextual) things considered, from communicative goals to interlocutor relationships and Enron culture. On the basis of feedback from his recipients (i.e. the intentioned perlocutionary force is arrived at with little resistance), it is likely that Nemec knows that that particular string, such as *please review and let(')s discuss* or *attached is a clean and redlined version*, is a useful and effective way of getting things done. Therefore, as subsequent potential

'usage events' (Langacker 2000: 9) arise in his day-to-day work, characterised by identical or similar contextual and communicative stimuli, he activates and uses the same "tried and tested" phrases. Kuiper (2004: 39) argues that 'formulaic performance' can only occur in communicative situations "where there is an expectation that things will happen in much the same way as they have happened before". It may be inferred from Nemec's recurrent and preferred use of these n-grams that both he and his recipient have this kind of expectation. Over time, with repeated use, these strings have become entrenched. As Schmid (2016: 21) states, word strings can come to be entrenched in such a way, and that "once entrenched, these routines are activated more quickly and with less effort and are therefore more likely to be repeated." While we cannot be definitive on whether corpus results reflect psychological reality (Schmitt et al. 2004, Durrant & Doherty 2010) the findings here indicate that this process may account for Nemec's habitual use of these word n-grams.

While this is interesting from a usage-based perspective, the main concern in this paper is the extent to which these word n-grams are idiolectal. For the idiolectal nature of 'entrenched' or even simply recurrent word strings to be proven, there needs to be evidence demonstrating that when faced with the same communicative situation, characterised by identical or similar contextual stimuli, different people produce different linguistic output. To do this, Nemec can be compared with the other 175 authors in the Enron corpus. The Enron corpus is relatively small, especially when being used as population data to make claims of linguistic uniqueness. However, emphasis in forensic linguistics is on the use of *relevant* population data (Grant 2013, Turell & Gavaldà 2013). While there are larger general reference corpora that could be used to find out how common or rare a linguistic feature is, they are not comparable across contextual variables such as time, text-type and medium, and the data of individual authors is not delineated. Therefore, more relevant and comparable corpora, though much smaller, are more useful. In this case, the remaining 175 authors of the Enron corpus arguably represent the most relevant population data possible; they are writing in the same medium, in the same community of practice, working for the same company at the same time and many have the same job as Nemec. Using the variant/variable paradigm (Mollin 2009: 382), we can observe the different ways in which authors express semantically and pragmatically the same thing. There are five

other authors in the corpus, all lawyers like Nemec, who also draw their recipients' attention to attached redlined versions of documents. However, none of them use the same *attached is a clean and redlined version* as Nemec (Table 6). Some refer to *redlined* versions but not *clean* (Dickson), some refer to *draft* instead of *version* (Jones), some use *redline* as a noun rather than the adjectival *redlined* (Mann, Perlingerie, Shackleton), and most use other ways of saying *attached is,* such as *I am attaching*, *here is* and *this is*. Shackleton comes the closest to using Nemec's string; she uses *attached are the clean and redlined versions*, referring in the plural to multiple documents. Although very close to Nemec, he retains his distinctiveness. The case is very similar for *please review and let(')s discuss*. Overall in the corpus, *review* is the third most frequent collocate of *please* (after *let* and *call*) with the two co-occurring 518 times and being used by 59 authors. Therefore, asking someone to review something is a very common practice in Enron emails, and there are twenty other authors in the corpus who, like Nemec, write *please review and* and then request some further communication. In the interests of space and readability, the most frequent users are displayed in Table 7. Rather than *let(')s discuss*, other authors request *comments* (Dickson, Hyvl, Sweet), direct their recipient to *call* them (Shakleton), and they commonly write *let me know* (Dickson, Hyvl, Sager). Again Shackleton comes closest to Nemec, with *please review and let's talk*. Once more, however, Nemec's string retains its distinctiveness.

The analysis of both of Nemec's n-grams show that, despite there being other authors who are doing the same thing communicatively in their emails, nobody produces identical output to that of Nemec. Therefore, we can argue that (*attached is*) *a clean and redlined version* and *please review and let(')s discuss* are idiolectal to Nemec; they are observable manifestations of his unique linguistic repertoire when compared against relevant population data. Furthermore, these idiolectal word n-grams were integral to the accurate attribution of his disputed email samples.

Table 6. Other lawyers drawing attention to attached documents

Author	Output
Stacy Dickson	*Thank you. <Message-ID: > **Attached is a redlined version** of*
	*Reimbursement Agreement. **Attached is a redlined version** which*
Tana Jones	***Attached is a clean and redlined draft** of the*
	***I have also attached a redlined draft** reflecting the*
	***I am enclosing a redlined draft** of the guaranty. Please let*
Kay Mann	***Here's a redline** against our previous draft. There are lot*
	*but **here's a redline** for you to look at. I picked up the*
	***Here's the current redline**, incorporating most of the*
	***Here's the Enron redline**: Too much paper! Kay*
	***Here's the clean version (no redline)** of the most recent*
	***The following is a redline version** of the execution version*
	*and **have attached a redline version**. Herman, I dealt with*
	***This is the most current version redlined** against a previous*
Debra Perlingerie	***Attached is a redline version** incorporating ENA and Cabo*
	*Tammi **attached is a redline version** of the draft with*
	*Regarding my phone message, **attached is a redline** reflecting*
	***Attached is their redline** for Greeley Gas Company. I know we*
	***Attached is the final redline** for Duke they are ready to*
	***I am attaching their redline** for you to see. Please note*
	***Please look at the attached redline**. If you are in agreement*
	***Please look at the attached redline from Greeley** regarding*
	***Here is the old redline version as well as the clean** for*
	***Here is a redline and clean version** of requested changes by*
	***This is the redline version**. Veronica please review the*
	***This is the current version without redline**. Debra*
Sara Shackleton	***Attached are the clean and redline versions**, respectively.*
	***Attached are the clean and redline versions** for the swap.*
	*All: **Attached are clean and redlined versions** of the swap*
	***Attached are clean and redline versions** of the latest*
	***Attached are my comments (clean and redline)** to the foregoi*
	***Attached is a redline** based upon the document generated by.*
	***Attached is my redline**. Most of the following people*
	***Attached is my redlined version** of Stacy's form. Do we need*
	*Jarrod: **attached is my prior redline** with suggested comments*
	*Diane: **attached is the redline redraft** for the KCS trades.*
	***Attached is the redline**. With the exception of numbers (cap,*
	*Per my voice mail, **attached is the redlined version**.*
	*Blair: **Attached below is a redline version** of the consent*
	*RBC and **am attaching a redline** of changes. Since ECC will be*
	***I have attached a redline** of the actual confirm (without*
	*Diane: **Please see the redline** where I have incorporated a*

Table 7. Other authors using *please review* + further communication

Author	Output
Stacy	***Please review and give me your comments.*** *Thanks, Stacy*
Dickson	***Please review and give me your comments.*** *Stacy*
	Please review and let me know if you have any comments.
	David, ***Please review and let me know what you think.*** *I would run*
Mark	***Please review and give me your comments*** *on the attached*
Haedicke	***Please review and give me your brief thought****s on this*
	Please review and give your thoughts to me *and Randy.*
	Richard, ***please review and comment*** *asap. Mark*
	Jeff, ***please review and let's talk about it****. Mark*
	Please review and set up a meeting *with me to establish a plan*
Dan Hyvl	***Please review and comment****. Thank you.*
	to new master ***Please review and comment****.*
	Please review and give me your comments*.*
	Please review and give me your comments. *I have talked with*
	Please review and let me have your comments*. If you want ENA*
	Please review and let me know *if this is what you wanted.*
	Please review and let me know *if you have any further changes*
Elizabeth	***Please review and call me*** *with any questions or comments.*
Sager	***Please review and call me*** *with any comments. I will review again*
	Please review and let me know *if you have any comments.*
	Please review and let me know *if you have any comments*
	Please review and let me know *if this appropriate and I have inc*
	Please review and let either Shari or I know *if you have any com*
	Please review and forward comments *to me early Thursday morning*
	Could you ***please review and let me know*** *what you think of this*
Sara	***Please review and call me.***
Shakleton	*Raptor equity share swaps so* ***please review and comment****. Thanks.*
	Please review and let's talk*. Thanks. SS*
	Please review and provide me with your comments.
Twanda	***Please review and provide me with any comments****. Rick will be in*
Sweet	***Please review and provide me with any comments*** *that you may have*
	Please review and provide me with any comments. *Thanks.*
	Please review and provide me with any comments *that you may have*
	Please review and provide me with any comments *that you may have*
	Please review and provide me with any comments*. Thanks.*
	Please review and provide me with any comments. *Thanks Richard*
	Please review and provide me with any comments *that you may have*
	Please review and provide me with your comments.

6. Conclusions and implications

This study set out to apply a corpus approach to a forensic problem. Concepts such as lexical priming (Hoey 2005) and formulaic language (Wray 2002) hold that individuals have their own unique store of collocations and word sequences, and a small amount of research from corpus linguistics (Mollin 2009, Barlow 2013) and authorship attribution (Hoover 2002, Juola 2013, Larner 2014) has demonstrated and exploited the idiolectal nature of word strings. This study aimed to build on this existing work, and to make a case for word n-grams as theoretically-motivated features for authorship analysis that can be used to attribute texts to their correct authors, and for which differences between authors can be explained.

Results of an authorship attribution experiment using word n-grams between 2 and 6 words long were promising, especially when attributing larger samples, where success rates reached as high as 95%. However, what was most notable in the results was that word n-grams captured some authors' idiolects better than others. In fact, different lengths of n-gram worked better for different authors. These differences in performance were apparently regardless of the amount of data available in the disputed samples of the authors, which indicates that size is not everything, and that it is the authors' particular style or idiolect which determines the success of an approach. This may motivate future authorship studies to more closely consider the success rate of methods on *individuals*, rather than making generalised claims about reliability and accuracy across *whole corpora*. After all, we are in the business of identifying individuals.

The case study analysis of Nemec focused on two distinctive word n-grams that were useful in identifying his disputed samples. It was suggested that these word n-grams were tied to recurrent and routine communicative situations that he encounters during his daily work. The relationship between word strings, phrases or 'formulae' and specific purposes is well-established (Becker 1975, Nattinger & DeCarrico 1992, Wray 2002, Kuiper 2004). By extension, the usage-based process of entrenchment (Langacker 1988, 2000; Schmid 2016) holds that on the basis of their unique socio-historical-linguistic characteristics, experiences and encounters, which word strings become entrenched inherently varies from author to author. Against this theoretical backdrop, it

was demonstrated that when faced with the same communicative situation and purpose, Nemec's linguistic output is different from other authors in the relevant population and why, ultimately, these word n-grams have worked in attributing his disputed samples.

For forensic linguists involved in developing methods for analysing authorship, the results and discussion presented here may provide an impetus for moving towards offering motivations for linguistic features used and explanations for results obtained in attribution experiments and casework. The linguistic features used in authorship analysis do not exist in a vacuum; rather, as demonstrated by word n-grams, they should be considered as observable, surface-level products of more complex pragmatic and communicative choices made by writers. As style markers, word n-grams require further testing on a wider range of less explicitly routine or formulaic text-types to evaluate their usefulness. For corpus linguistics, this study contributes to other work (Coniam 2004, Mollin 2009, Barlow 2013) in demonstrating the opportunities and benefits of using specialised corpora to produce evidence supporting the theory of idiolect. Overall, a desirable outcome would be for researchers to more regularly traverse the interface between corpus linguistics and forensic authorship attribution first explored by Svartvik (1968) and Coulthard (1994) decades ago.

Acknowledgments

This research was supported by the Arts and Humanities Research Council (AHRC) grant reference AH/J502686/1. I sincerely thank the two anonymous reviewers of this paper for their very helpful and constructive comments.

References

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

Barlow, M. (2013). Individual differences and usage-based grammar. *International Journal of Corpus Linguistics, 18*(4), 443–478.

Becker, J. D. (1975). The phrasal lexicon. In B. L. Nash-Webber & R. Shank (Eds.), *Theoretical Issues in Natural Language Processing* (pp. 60–63). Cambridge, MA: Bolt Beranek and Newman.

Biber, D., Conrad, S., & Cortes, V. (2004). *If you look at…*: Lexical bundles in university teaching and textbooks. *Applied Linguistics, 25*(3), 371–405.

Bloch, B. (1948). A set of postulates for phonemic analysis. *Language, 24*(1), 3–46.

Cohen, W. W. (2009). *Enron Email Dataset* [online]. Retrieved from http://www.cs.cmu.edu/~enron/ (last accessed November 2010).

Coniam, D. (2004). Concordancing oneself: Constructing individual textual profiles. *International Journal of Corpus Linguistics, 9*(2), 271–298.

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

Coulthard, M. (1994). On the use of corpora in the analysis of forensic texts. *Forensic Linguistics. 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., Grant, T., & Kredens, K. (2011). Forensic Linguistics. In R. Wodak, B. Johnstone & P. Kerswill (Eds.), *The SAGE Handbook of Sociolinguistics* (pp. 531–544). London: Sage.

Coyotl-Morales, R., Villaseñor-Pineda, M. L., Montes-y-Gómez, M., & Rosso, P. (2006). Authorship attribution using word sequences. In J. F. Martínez-Trinidad, J. A. Carrasco Ochoa & J. Kittler (Eds.), *Proceedings of the 11th Iberoamerican Congress on Pattern Recognition* (pp. 844–853). Berlin: Springer.

Durrant, P., & Doherty, A. (2010). Are high-frequency collocations psychologically real? Investigating the thesis of collocational priming. *Corpus Linguistics and Linguistic Theory*, *6*(2), 125–155.

Eckert, P., & McConnell-Ginet, S. (1998). Communities of practice: Where language, gender and power all live? In J. Coates (Ed.), *Language and Gender: A Reader* (pp. 484–494). Oxford: Blackwell.

Eder, M. (2015). Does size matter? Authorship attribution, small samples, big problem. *Digital Scholarship in the Humanities, 30*(2), 167–182.

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

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 & M. T. Turell (Eds.), *Dimensions of Forensic Linguistics* (pp. 215–229). Amsterdam: John Benjamins.

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.

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

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.

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

Juola, P. (2008). *Authorship Attribution*. Delft: NOW Publishing.

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

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

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* (pp. 405-437). Peter Lang: Frankfurt am Mein.

Kuiper, K. (2004). Formulaic performance in conventionalised varieties of speech. In N. Schmitt (Ed.), *Formulaic Sequences: Acquisition, Processing and Use* (pp. 37–54). Amsterdam: John Benjamins.

Langacker, R. (1988). A usage-based model. In B. Rudzka-Ostyn (Ed.), *Topics in Cognitive Linguistics* (pp. 127-161). Amsterdam: John Benjamins.

Langacker, R. (2000). A dynamic usage-based model. In M. Barlow & S. Kemmer (Eds.), *Usage-Based Models of Language* (pp. 1-63). Stanford: CSLI Publications.

Larner, S. (2014). A preliminary investigation into the use of fixed formulaic sequences as a marker of authorship. *International Journal of Speech, Language and the Law, 21*(1), 1–22.

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

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

Mikros, G. (2012) Authorship attribution and gender identification in Greek blogs. In I. Obradovic, E. Kelih & Reinhard Köhler (Eds.), *Methods and Applications of Quantitative Linguistics* (pp. 21–32). University of Belgrade: Academic Mind.

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

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

Nini, A., & Grant, T. (2013). Bridging the gap between stylistic and cognitive approaches to authorship analysis using Systemic Functional Linguistics and multidimensional analysis. *International Journal of Speech, Language and the Law, 20*(2), 173–202.

Sanderson, C., & 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* (pp. 482–491). Morristown, NJ: Association for Computational Linguistics.

Schmid, H-J. (2016). A framework for understanding linguistic entrenchment and its psychological foundations. In H-J. Schmid (Ed.), *Entrenchment and the Psychology of Language Learning: How We Reorganize and Adapt Linguistic Knowledge* (pp. 9–36). Berlin: De Gruyter Mouton.

Schmitt, N., Grandage, S., & Adolphs, S. (2004). Are corpus-derived recurrent clusters psycholinguistically valid? In N. Schmitt (Ed.) *Formulaic Sequences: Acquisition, Processing and Use* (pp. 12 –151). Amsterdam: John Benjamins.

Scott, M. (2008). *WordSmith Tool* (Version 5) [Computer software]. Liverpool: Lexical Analysis Software.

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

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.

Svartvik, J. (1968). *The Evans Statements: A case for Forensic Linguistics*. Gotëborg: University of Gothenburg Press.

Turell, M. T., & Gavaldà, N. (2013). Towards an index of idiolectal similitude (or distance) in forensic authorship analysis. *Journal of Law and Policy, 21*(2), 495–514.

Woolls, D. (2013). *CFL Jaccard n-gram Lexical Evaluator* (*Jangle*) (Version 2) [Computer software]. CFL Software Limited. Retrieved from http://www.cflsoftware.com/ (last accessed January 2017).

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. (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.

Author's address

David Wright
School of Arts and Humanities
Nottingham Trent University
Clifton
NG11 8NS
United Kingdom

david.wright@ntu.ac.uk