An authorship analysis of the Jack the Ripper letters

DOI:
10.1093/llc/fqx065

Document Version
Accepted author manuscript

Link to publication record in Manchester Research Explorer

Citation for published version (APA):

Published in:
Digital Scholarship in the Humanities

Citing this paper
Please note that where the full-text provided on Manchester Research Explorer is the Author Accepted Manuscript or Proof version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version.

General rights
Copyright and moral rights for the publications made accessible in the Research Explorer are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Takedown policy
If you believe that this document breaches copyright please refer to the University of Manchester’s Takedown Procedures [http://man.ac.uk/04Y6Bo] or contact uml.scholarlycommunications@manchester.ac.uk providing relevant details, so we can investigate your claim.
1. An authorship analysis of the Jack the Ripper letters

Dr Andrea Nini

andrea.nini@manchester.ac.uk

Samuel Alexander Building
The University of Manchester
Oxford Rd
Manchester
M13 9PL
UK
1.1. Introduction

On the 31st of August 1888, the murder of a prostitute in the Whitechapel area of London started a series of homicides that would be remembered for over a century: the Whitechapel murders. These murders were characterized by mutilations of increasing gruesomeness, such as disembowelment or removal of organs. Experts believe that between five and six murders were committed, culminating with the most violent one on the 9th of November 1888. Although the killings are traditionally attributed to a single person, commonly known as Jack the Ripper, there has never been definitive evidence to exclude the possibility that the murders were unconnected events, despite some modern research that has found the set of shared behavioural characteristics of the murders to be distinctive (Keppel et al., 2005).

Besides the investigative aspect, the Whitechapel murders case and the legend of Jack the Ripper have an important socio-cultural dimension. The mystery surrounding the identity of the killer has led to incredible and often unlikely speculations and even though the Whitechapel murders happened more than a century ago, the mystery has created a business that is still alive and generating revenue in the form of media products, books, and tours. These elements have contributed to the engraving of the mythology of Jack the Ripper into modern Western culture far more than the murders themselves and several academic works have explored both the sociological dimension of the mythology of Jack the Ripper to shed light on 19th century England and the beginning of the modern era (Walkowitz, 1982; Perry Curtis, 2001; Haggard, 2007) or have identified links between Jack the Ripper and Victorian literature (Tropp, 1999; Eighteen-Bisang, 2005; Storey, 2012).

The origin of the mythology of Jack the Ripper lies in the communication that the killer allegedly sent to the police or media during the time of the murders and in the following months and years. Although there is no evidence that the real killer was involved in the production of any of them, the more than 200 Jack the Ripper letters significantly contributed to the creation and popularisation of the name and persona of Jack the Ripper. However, despite the large number of
texts involved in the case, only a small number of the Jack the Ripper letters received substantial investigative or socio-cultural importance at the time.

Probably the most important text in the case is the *Dear Boss* letter, which was received on the 27th of September 1888 by the Central News Agency of London. This letter is the first ever signed as 'Jack the Ripper' and it is responsible for the creation of the pseudonym. The letter claimed responsibility for the murder of Annie Chapman on the 8th of September 1888 and mentioned that an ear would be cut off from the next victim and sent to the police. Indeed, the murder of Chapman was followed by another murder in which part of one of the ears of the victim was removed, although this was never sent to the police. Because of this fact and its style and content, the letter was considered to be genuine and it became famous for introducing the persona of Jack the Ripper and for providing a name that the press could use to refer to the killer.

The second most important text is the *Saucy Jacky* postcard, which was received on the 1st of October 1888 by the Central News Agency of London, signed again as 'Jack the Ripper'. The postcard claimed responsibility for the double murder of Elizabeth Stride and Catherine Eddowes on the night of the 30th of September 1888. The postcard did not threaten future murders and presented an apology for not having sent an ear to the police. Together with the *Dear Boss* letter, this postcard has also become iconic in the portrayal of Jack the Ripper and was taken more seriously than other letters because of the short window between the murders and the time the postcard was sent (Begg, 2004).

The police took these two texts seriously enough to produce and post copies outside of police stations on the 3rd of October 1888 (Rumbelow, 1979; Sugden, 2002). Following that, on the 4th of October, the two texts were also published in many newspapers (Sugden, 2002), even though some newspapers had obtained the information of the name 'Jack the Ripper' and part of the texts already by the 1st of October (Perry Curtis, 2001).

Although much less popular than the other two texts, on the 5th of October the Central News Agency also received a third text, commonly known by experts as the *Moab and Midian* letter. This
text announced a triple event and justified the murders with religious motives. The peculiarity of this letter is that the original had never been sent to the police, as the journalist Tom Bulling of the Central News Agency decided to copy the text and send only the envelope to the police. The reasons behind this choice were not explained and to date they are still unknown.

Besides the three texts delivered to the Central News Agency, a large number of other letters and postcards were sent to several other recipients such as the press or the police between October 1888 and November 1888, that is, after the two iconic texts were made public by the police. During this period, 130 letters allegedly written by the killer were received, and the flow of letters continued for ten more years. Among these letters, another text that has become iconic and that was judged as important during the case is the *From Hell* letter, which was received on the 16th of October by George Lusk, head of the Whitechapel Vigilance Committee, together with half of a kidney (Rumbelow, 1979).

In most of the letters, the author(s) mimicked the original *Dear Boss* letter and *Saucy Jacky* postcard in terms of taunting the police and using salient stylistic features, such as the laughter *ha ha*, or the salutation *Dear Boss*. Some of the letters were almost exact copies of *Dear Boss*, especially the ones that were received a year or more later, in conjunction with the anniversary of the murders or in conjunction with new murders in Whitechapel.

Since it is quite unlikely that the same person produced hundreds of letters spanning decades and sent from different places across the UK, it is commonly assumed that most of the letters were written by different individuals, who possibly had not been involved with any of the killings. Particularly interesting is the case of Maria Coroner, a 21 year old girl who was caught sending one of those letters (Evans and Skinner, 2001). When questioned, she explained that she did so as she was fascinated by the case. It is likely that many of the writers of these letters acted for similar reasons, although the motives behind such actions will probably never be established. These hoax letters themselves represent an interesting mirror into the fears and problems of the people who wrote them (Remington, 2004). More importantly, these letters still exercise an impact on modern
times. The Yorkshire ripper hoaxter, for example, sent letters that borrowed several linguistic elements from the *Dear Boss* letter (Ellis, 1994; Lewis, 1994).

Such a collection of letters also represents an invaluable data set for forensic linguistics and for authorship analysis. Linguistic analyses of the letters can be useful to provide new evidence for the Whitechapel murders case, since, as opposed to other sources of evidence nowadays corrupted by time, the language of the letters has reached us unchanged. The question of the authorship of the letters mostly focuses on the early ones, such as the *Dear Boss* and *Saucy Jacky* texts. The most common theory about the authorship of these texts is that journalists fabricated them in order to increase newspaper sales. The 'enterprising journalist' theory, more specifically, suggests that letters such as the *Dear Boss* letter were actually works of fiction skilfully created to generate shock and 'keep the business alive' (Begg, 2004; Begg and Bennett, 2013). Evidence for the 'enterprising journalist' theory comes from the *Littlechild* letter, in which Detective Chief Inspector John George Littlechild mentions that at Scotland Yard virtually everyone knew that the *Dear Boss* letter was fabricated by Tom Bulling, a journalist of the Central News Agency itself, in collaboration with his manager (Rumbelow, 1979; Begg, 2004). At the time, the Central News Agency had been in a fierce competition with other news agencies and had a reputation of fabricating or embellishing news (Evans and Skinner, 2001; Begg, 2004). Another theory proposed by Cook (2009) suggests that a journalist named Frederick Best from the tabloid newspaper *The Star* was the actual author of the *Dear Boss* letter.

As a first step to shed light on the authorship question of the Jack the Ripper letters, the present paper reports on an authorship analysis of the texts received during and after the Whitechapel murders case that are connected to Jack the Ripper. The available data set lends itself to several authorship questions, such as the profiling of the anonymous author(s), or to the comparison between some key letters and Bulling's and Best's writings. In the present paper an initial exploration of the Jack the Ripper letters is performed with the general aim of finding out for which of the hundreds of texts there is evidence of common authorship, with a special attention to
the most important texts in the case mentioned above and on those earliest texts received before the
1st of October 1888, that is, before the Dear Boss letter and the Saucy Jacky postcard became of
public domain.

Establishing whether some of the Jack the Ripper texts could be written by the same person
is an important preliminary step as any future study, either involving profiling or comparison,
would benefit from knowing if a number of questioned texts can be clustered together. In this sense,
the authorship question tackled in the present study constitutes a useful starting point for any future
authorship study on the Jack the Ripper letters.

1.2. Data

The data set used in the present study is a corpus that includes the texts connected to the
Whitechapel murders: the Jack the Ripper Corpus (JRC). This corpus consists of the letters or
postcards found and transcribed in the Appendix of Evans & Skinner (2001), who claim to have
collected all of the texts involved in the Whitechapel murders related to Jack the Ripper from the
Metropolitan Police files. These letters were OCR-scanned from the book and the scans were
manually checked for scanning errors. The corpus consists of 209 texts and 17,463 word tokens.
The average length of a text in the corpus is of 83 tokens (min = 7, max = 648, SD = 67.4).

The peculiarity of the JRC is that almost all of the texts in the corpus are comparable in
terms of their broad situational parameters (Biber, 1994), as they are almost all written letters or
postcards with similar linguistic purposes. For example, in terms of addressee, 67% of the texts
were addressed to Scotland Yard; Sir Charles Warren, the head of London Metropolitan Police
during that time; Inspector Abberline; or other law enforcement units. The remaining 33% were
either of unknown addressee (13%), or were addressed to common citizens or to newspapers, news
agencies, schools, or private firms (20%). The vast majority of the letters was postmarked or found
in London, although other letters were postmarked or found in places all over the UK, such as
Birmingham, Bradford, Dublin, Edinburgh, Liverpool, Manchester, or Plymouth. All of the letters
were handwritten and a minority of them (4%) included drawings of various items, such as knives,
skulls, or coffins. Finally, a large number of the letters (75%) were indeed signed as 'Jack the Ripper' or with variants of the name, such as 'Jack the Whitechapel Ripper', or 'JR', or 'jack ripper and son'. Some other letters were not signed (11%) while the remaining letters used other pseudonyms, such as 'Jim the Cutter', 'The Whore Killer', or 'Bill the Boweler'.

The corpus ranges from the 24th of September 1888 to the 14th of October 1896, thus spanning more than 10 years after the murders. However, the majority of the texts, that is 62% of the corpus, was received during the period between October 1888 and November 1888.

Among the total set of 209 texts, the present analysis will pay special attention to those early texts that were received not later than the 1st of October 1888, before the content of the Dear Boss and Saucy Jacky was popularised by the police and the media and therefore hoaxers could have knowledge of it. Before this date, according to Evans & Skinner's (2001) collection, four texts were received:

- Text 1 (24th September, 128 word tokens): In this text the author admits to the killing of Chapman and presents the intention to stop killing. The letter is unsigned.
- Text 2 (27th September, 244 word tokens): The Dear Boss letter.
- Text 3 (1st October, 57 word tokens): The Saucy Jacky postcard.
- Text 4 (1st October, 88 word tokens): This text threatens more murders and is signed as 'Ripper'.

Even though the analysis will include all the JRC texts, these four texts are particularly important because any linguistic similarity that links them cannot be explained by influence from the media, an explanation that cannot be ruled out for the other texts. In the rest of this paper, the four texts above will be called the pre-publication texts, whereas the remaining 205 texts will be called the post-publication texts.

1.3. Methodology

The authorship question considered for this study concerns finding out which texts in a corpus are likely to be written by the same author. Recently, this task has been called author
clustering and it has been tackled using hierarchical cluster analysis on frequencies of features (Gómez-Adorno et al., 2017). This authorship problem could be considered, however, just as a special case of authorship verification, a problem that has received considerable attention in the literature (Koppel and Schler, 2004; Koppel et al., 2012; Brocardo et al., 2013; Koppel, Schler and Argamon, 2013). The best solutions proposed to solve this type of problem involve the addition of distractor texts belonging to similar registers and the use of similarity metrics applied to feature sets consisting of frequencies of linguistic features.

The problem in applying any of these techniques to the JRC corpus is that the JRC texts are too short to produce reliable frequencies, as the average text length for the corpus is only 83 word tokens. For this reason, in this case it is necessary to adopt a method that does not involve the computation of frequencies.

A solution to the problem of analysing short texts within a forensic linguistic context by considering the presence or absence of features as opposed to their frequencies has been initially proposed by Grant (2010) and then further described in Grant (2013) for text messages. Inspired by research in similarity between species in biology and ecology, and already applied to assess similarity in crime types, this approach consists in quantifying the similarity between two texts using the Jaccard coefficient, or the number of shared features between two texts divided by the total number of features in both texts (Jaccard, 1912):

\[ J(A,B) = \frac{|A \cap B|}{|A \cup B|} \]

After being successfully applied to text messages case, methods using the Jaccard coefficient have been applied with good results to other registers, including newspaper articles (Juola, 2013), short emails (Johnson and Wright, 2014; Wright, 2017), and elicited personal narratives (Larner, 2014). These studies have analysed the presence/absence of combination of words, mostly looking at word n-grams, that is, strings of words of length \( n \) collected using a moving window.

Within plagiarism detection research, word n-gram techniques based on similar
mathematical principles are very common (Oakes, 2014: 65) on the grounds that the more shared strings there are in two documents, the more there is shared similarity of encoding of meanings and therefore the less likely it is that the documents are independent from each other, as explained by Coulthard (2004).

Word n-grams have been extensively adopted as linguistic features in traditional frequency-based stylometric methods for authorship attribution, although they are not deemed the best stylometric features, as they are often surpassed in efficacy by function words, simple word frequency, and, above all, character n-grams (Grieve, 2007; Stamatatos, 2009). Although word n-grams might not be extremely good features when frequency is taken under consideration, for a method involving presence/absence these features are much better than single words or function words because word strings are rarer and the power of a presence/absence method lies in the measurement and comparison of the linguistic uniqueness of each author on rare features. Character n-grams could also be good features but they are less amenable to interpretation, which can be a drawback depending on the ultimate goal of the research.

In addition to these methodological advantages, the use of word n-grams as features has theoretical support. Corpus linguistics (Sinclair, 1991; Biber, Conrad and Cortes, 2004; Hoey, 2005) and psycholinguistics/cognitive linguistics (Langacker, 1987; Barlow and Kemmer, 2000; Schmitt, 2004; Wray, 2005; Schmid, 2016) have long theorised that combination of words are at the core of language processing and empirical support has been found for these theories (Ellis and Simpson-Vlach, 2009; Tremblay, Derwing and Libben, 2009).

Furthermore, there is also empirical support for a strong idiolectal effect in the production and processing of word combinations (Mollin, 2009; Barlow, 2013; Schmid and Mantlik, 2015; Günther, 2016). Wright, (2017) reveals the idiolectal nature of certain word n-grams by taking one specific speech act as constant and then analysing how different authors realise this act, uncovering that each author recurs to their own idiosyncratic set of lexical choices to perform the same act.

In the present study, for the reasons explained above, the set of features that is taken under
consideration is word n-grams, as the ultimate goal is to discover possible idiolectal encoding in the
JRC letters. Because the JRC texts are short, presence or absence of word n-grams is considered, as opposed to their frequency. Among all the possible sizes of n-grams, word 2-grams are chosen as any n-gram of \( n > 2 \) is ultimately made up of n-grams of \( n = 2 \), meaning that word 2-grams return the most complete picture of the shared word combinations in two sets. Presence or absence of word n-grams is quantified using the Jaccard distance, as opposed to the coefficient, which can be defined as:

\[
d_J(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|}
\]

and which returns values between 0, or absolute identity, and 1, or absolute distance. The Jaccard distance is used so that a hierarchical cluster analysis can then be carried out. In this way, it is possible to firstly find out the major groups of texts that are more similar to each other, and then it is possible to zoom in and explore smaller groups of letters, such as the pre-publication letters.

However, evidence of common authorship of two sets of documents can come not only from finding similarity but also from establishing that this similarity is distinctive (Grant, 2010, 2013). Although it is difficult to establish a universal threshold for distinctiveness, it is safe to assume that if a particular n-gram or lexicogrammatical structure does not occur at all or occurs extremely infrequently in a comparable reference corpus then this n-gram or structure is distinctive.

The comparison corpus used to assess distinctiveness should therefore include relevant population data (Turell and Gavaldà, 2013; Wright, 2017). If a smaller sub-sample of its texts is considered, the remaining of the JRC itself is indeed a corpus with relevant population data. However, because of its relatively small size, more data from 19th century English is necessary to find evidence of distinctiveness. Ideally, because of the pervasiveness of register variation, the perfect comparison corpus would be one including a large number of 19th century English letters of comparable communicative situation (Biber, 2012). However, in absence of an extensive resource of this kind, the most comprehensive largest available set of general reference corpora was used instead, consisting of the largest available corpora of 19th century English:
• The 132 million word 19th century section of the Corpus of Historical American English (COHA);
• The 34 million word Corpus of Late Modern English Texts 3 (CLMET3), spanning from 1710 to 1920;
• The 19 million word Extended Old Bailey Corpus (EOBC), including the proceedings of the Old Bailey from 1720 to 1913.

In sum, the method adopted in this study involves the comparison of all the texts in the JRC to each other using the Jaccard distance and a set of comparison corpora in order to find whether there are texts that are similar and distinctive in their linguistic encoding.

In addition, since the analysis involves word n-gram types, the method faces problems when dealing with texts of different length, as the likelihood of any word or n-gram type being observed is correlated with text length. However, provided that the shared n-grams found are also highly distinctive the evidence of common authorship is nonetheless valid despite differences in text lengths.

1.4. Results

Figure 1 reveals that the relationship between the percentage of texts using a 2-gram (occurring in at least 2 texts) and their frequency rank form a zipfian shape, as expected (Zipf, 1935). The graph shows that the top eight 2-grams appear in at least 20% of the corpus. Some of these are very frequent because they reflect common grammatical structures of English, such as I am, I have, I will. Two 2-grams reflect the influence of the signature and salutation of the Dear Boss letter on the rest of the corpus: jack the and dear boss. Finally, the high incidence of the 2-grams I shall and yours truly are probably explained by both the influence of the Dear Boss letter and by the register of the letters.
Because of their frequent occurrence and thus reduced discriminatory power, these top eight 2-grams were excluded from further analysis.

The distance between each pair of texts was quantified using the Jaccard distance based on the presence or absence of the remaining 1541 word 2-grams and a distance matrix was therefore generated. Figure 2 shows a histogram and boxplot of the Jaccard distances for all possible pairs of texts in the JRC.
Figure 2 - Histogram and boxplot showing the distribution of Jaccard distance values for all possible pairs of texts in the JRC

As the histogram of Figure 2 shows, the most frequent Jaccard distance and also the median distance is approximately 1, which generally speaking means that the texts in the JRC are not very similar to each other. Only 25% of the scores are lower than 0.98, which is marked in Figure 2 by the leftmost edge of the boxplot, and only 6% of the scores are lower than 0.95, that is, the outliers in the boxplot of Figure 2 indicated by circles.

The distance matrix was then used for a hierarchical cluster analysis that can be visualised through the radial dendogram in Figure 3.
Figure 3 - Radial dendogram displaying the results of a hierarchical cluster analysis of the JRC corpus using the Ward method based on Jaccard distances. The name of the texts is a code starting with two letters from the signature and followed by the date in which it was received. The texts mentioned in the introduction, including the pre-publication texts, contain their name in addition to the code.

Three main branches stem from the centre of the graph in Figure 3, corresponding to the three main clusters found. On the right, there are two main clusters, one of which includes only two texts. The remaining texts are all classified into another cluster whose branch points to the left and that further splits into two other clusters that roughly correspond to the two hemispheres of the graph. The most historically interesting texts, including the pre-publication texts, are all grouped in the cluster spanning over the top hemisphere of the graph and therefore the rest of the paper will focus on this cluster. Although it would be interesting to explore the other clusters, this is beyond the scope and space of this study. The branch leading to the top hemisphere then splits even further into two more sub-branches, one developing to the right containing the From Hell letter, and one to the left where
all the other historically important letters, including the pre-publication texts, are grouped. The split
at this level suggests that the From Hell letter is rather linguistically dissimilar to the other famous
letters, at least in terms of word 2-gram use. The left branch then splits into two more clusters, with
the rightmost one splitting again into two large clusters. One of these two contains the Dear Boss
letter, the Saucy Jacky postcard, and the Moab and Midian letter, while the one next to it contains
two of the pre-publication letters. Therefore, among the pre-publication JRC texts, Dear Boss and
Saucy Jacky are the most similar one, with the Moab and Midian letter being the most similar to
them among all the historically important texts.

1.4.1. The pre-publication texts

Let us therefore examine the pre-publication texts using a network graph as in Figure 4, in
which each circle represents a text with a size proportional to the text’s length in total word tokens
and each link represents an overlapping word 2-gram type. The graph also reports the Jaccard
distance for each pair of texts.
As the cluster analysis already suggested, it is evident that the two pre-publication texts that are more similar to each other are the Dear Boss letter and the Saucy Jacky postcard. Additionally, these two texts have a Jaccard distance of 0.93, which is a degree of similarity that can be found in less than 5% of the pairs of texts in the JRC. The amount of shared language is striking considering the fact that the Saucy Jacky postcard is very short and does not share any linguistic link with either the 24th September text or the 1st October text. Although the Dear Boss letter shares a number of 2-grams with both Text 1 and Text 4, the Jaccard score for both pairs is in the average for the corpus.

Excluding the 3-gram Jack the Ripper, which refers to the signatures of the two texts, Table 1 below presents the concordances of their overlapping 2-grams, with an analysis of their syntactic structure.
A closer examination reveals that the two texts share 2-grams of varying distinctiveness. The phrase *a bit*, although with different syntactic function (1), the verbs *give* (2) and *got* (3), or the use of the infinitive verb *to get* (3) are common structures that are frequently found both in the JRC and in reference corpora of 19th century English. The use of *till* as a variant of *until* is also not very distinctive as it is the predominant variant in the JRC (80%), CLMET3 (75%), and EOBC (90%) but not in 19th century COHA (28%).

The two texts also share the use of infinitive clauses to post-modify the noun *time* with a negation in the matrix clause (6), which occurs in only two other texts in the JRC. The structure is quite rare even at a more general level, as it is found about 10-18 times per million words across the reference corpora.

*Dear Boss and Saucy Jacky* also share the use of the verb *work* to euphemistically indicate the act of killing (7). This use of *work* is found in some post-publication JRC texts (about 20% of
the texts in the corpus. It is very difficult to estimate distinctiveness for (7) using larger reference 
corpora, however, as it would involve the manual analysis of thousands of instances.

Finally, the two texts share the use of a verb phrase headed by the phrasal verb to keep back 
with the direct object being a noun phrase with letter as head followed by a subordinate clause 
introduced by the subordinator till (8). Indeed, the 4-gram letter back till I itself neither occurs in 
any other text in the JRC, nor in any other corpus of 19th century English listed above. Because of 
the rarity of this 4-gram and the absence of a more relevant large corpus of 19th century English 
letters, a search of this 4-gram was performed on the web, which returned a total of 2,640 hits, all 
from exact copies of either the Dear Boss letter or of the Saucy Jacky postcard.

A search was then performed using the 19th century English corpora listed above on the 
overall phrasal verb keep back used with the meaning of withholding a letter as opposed to the use 
of the other synonyms keep (without back), hold back, hold up, hold out, withhold, delay plus any 
verb indicating ‘sending’, refrain plus any verb indicating ‘sending’, and detain. The queries were 
carried out by searching for occurrences of the lemma of these variants accompanied by the lemma 
LETTER within a span of ± 7 words. The concordances were then manually examined to count only 
instances of the meaning of 'withholding a letter, delay a letter to be sent'.

This corpus search revealed that keep back is found 22.5% of the times the meaning of 
‘delay sending a letter’ is expressed across the reference corpora. In the 19th century section of 
COHA, the majority of the instances (59%) use the variant withhold. Out of seven instances of keep 
back, four are from one author, John Townsend Trowbridge. In CLMET3 the most common variant 
is again withhold. One judge also uses keep back in the EOBC, where the most common variant is 
instead detain. Finally, in the JRC, only three instances of this meaning are found, two of which are 
instances of keep (quiet) found in two letters dated, respectively, 20 October (‘keep this quiret [sic] 
till I have done one’) and 09 November 1888 (‘keep this letter a bit quiert [sic] till you here of me 
again’). The third one is found in the Moab and Midian letter and it is the only instance across all 
the corpora to exactly match the syntactic structure in (8), Error! Reference source not
found. having the object in between the main verb and the particle as well as a subordinate clause introduced by the subordinator till (‘keep this back till three are wiped out’).

In conclusion, among the four pre-publication texts, these results support the hypothesis that the Dear Boss and Saucy Jacky texts were not written independently from each other, since these two texts are more similar to each other in their use of word 2-grams than 95% of all the other possible pairs of texts in the JRC even though the texts received later could have been influenced by them, and since some of these similarities are also distinctive.

1.4.2. The post-publication texts

Having established a link between the Dear Boss letter and the Saucy Jacky postcard, let us now explore the post-publication texts to determine whether further links between these two texts and other texts can be found.

As Figure 5 indicates, only eight JRC texts have a Jaccard distance lower than 0.95 with Dear Boss, including Saucy Jacky ($d_J = 0.929$) and Moab and Midian ($d_J = 0.934$), which are both therefore more similar to Dear Boss than 95% of the JRC. The most similar text to Dear Boss is, however, JR_191188, with a Jaccard distance of 0.776. This is not reported in Figure 5 to ease the visualisation of the boxplots.
Figure 5 - Boxplots showing Jaccard scores for Dear Boss (left) and Saucy Jacky (right) and all the other texts in the JRC

However, this text can be discounted as its anomalous score is explained by the fact that most of it was copied verbatim from Dear Boss, as the presence of an overlapping 13-gram demonstrates:

_I want to get to work right away if I get a chance_ and will do another one indoors. (JR_191188)

My knife’s so nice and sharp _I want to get to work right away if I get a chance_. (Dear Boss)
This is somewhat expected in the post-publication texts as the *Dear Boss* and *Saucy Jacky* were in the public domain.

For *Saucy Jacky*, Figure 5 indicates that the median score is 1 and that 50% of the texts in the JRC therefore have almost no linguistic link with it. Only 12 JRC texts have a Jaccard distance lower than 0.96, and, among these, the *Moab and Midian* letter is even more striking as its Jaccard score with *Saucy Jacky* is 0.90, which is 0.03 points smaller than the second most similar text, the *Dear Boss* letter.

From this analysis it is evident that the *Moab and Midian* letter not only is the most similar to *Saucy Jacky* of all the other texts in the JRC, but it is also almost as close as *Dear Boss* is to *Saucy Jacky* and, more importantly, it is the only text that is very close to both *Saucy Jacky* and *Dear Boss* (with the exclusion of the JR_191188 that contains a 13-gram copied word-by-word).
till I do [NP a bit more work] (Dear Boss)
number one squealed [ADV a bit] (Saucy Jacky)
will send you [NP a bit of face] by post (Midian)

I love [NP my work] (Dear Boss)
The police now reckon [NP my work] a practical joke (Midian)
you'll hear about [NP [NP saucy Jacky] [Gen s] [N work]] (Saucy Jacky)
well well [NP Jacky] [VP's [NP a very practical joker]] (Midian)

I keep on hearing [NP the police] have caught me (Dear Boss)
and send to [NP the police officers] (Dear Boss)
[NP The police] now reckon (Midian)

[NP [ADJ Grand] [N work]] the last job was (Dear Boss)
helps me in my [NP [ADJ grand] [N work]] (Midian)
is fit enough I hope [INTJ ha, ha.] (Dear Boss)
They say I'm a doctor now [INTJ ha ha] (Dear Boss)
Jacky's a very practical joker [INTJ ha ha] (Midian)

I wasn't coddling [NP dear old Boss] (Saucy Jacky)
I promise this [NP dear old Boss] (Midian)

thanks for [VP keeping [NP last letter] [PART back] [SUBCL till I got to work]] (Saucy Jacky)
[VP Keep [NP this] [PART back] [SUBCL till three are wiped out]] (Midian)

[Saucy Jacky's work] [ADVP tomorrow] [NP double event] [NP this time] (Saucy Jacky)
[Saucy Jacky's work] [ADVP tomorrow] [NP treble event] [NP this time] (Midian)

Table 2 - Syntactic analysis of the concordances for the n-grams in common between Dear Boss and Saucy Jacky and the Moab and Midian letters

Table 2 presents the 2-grams and their underlying syntactic structures shared by Moab and Midian and either Dear Boss or Saucy Jacky. Midian shares with the two pre-publication texts as well as with several other JRC texts the use of the phrase a bit (1) and the verb work to euphemistically mean kill (2). Midian and Saucy Jacky also share the use of the pseudonym Jacky, although the 2-gram Jacky's is only a surface similarity as its underlying syntactic structure is very different (3). Midian also presents the use of a pro-verb do (4) and it mentions the police (5), similarly to Dear
The adjective *grand* to modify *work* (6), the interjection *ha ha* (7), and the vocative *dear old boss* (8) are features that have been copied by other authors of the JRC texts, as they appear in, respectively, 3, 8, and 55 other JRC texts.

The two most distinctive structures are the verb phrase headed by *keep back* (9), already discussed above, and the use of a verbless clause, *ADJ–ble event this time*, elaborating the previous clause ending with the adverb *tomorrow* (10). This last syntactic structure is underlying the 2-gram *work tomorrow* and the 3-gram *event this time*, which do not appear in any other JRC text.

The 2-gram *work tomorrow* is surprisingly infrequent in the reference corpora (0.03–0.05 per million words) while the 3-gram *event this time* cannot be found at all. Although the 3-gram can be found on the web (617,000 hits), a search of the two n-grams together returns almost only instances of either *Saucy Jacky* or *Moab and Midian*.

In conclusion, there is linguistic evidence in support of the hypothesis that the *Moab and Midian* letter has an authorship link with the other two pre-publication texts, even accounting for the fact that *Dear Boss* and *Saucy Jacky* were publicly available at the time *Midian* was received.

### 1.5. Discussion

The analysis of the n-gram types reported above suggests that the *Dear Boss* letter and the *Saucy Jacky* postcard share distinctive linguistic similarities. Because authorship analysis studies demonstrated that common strings or rare collocations shared by documents are indicative of a common authorial source (Coulthard, 2004; Mollin, 2009; Johnson and Wright, 2014), given that the *Dear Boss* letter was not made public before the *Saucy Jacky* postcard was sent, the degree of their shared linguistic encoding is highly suggestive of the two documents not being produced independently. Although it is entirely possible that one author was responsible for all of the earlier texts, the linguistic evidence found so far can only suggest a link between the *Dear Boss* letter and the *Saucy Jacky* postcard while no strong links can be found between these two texts and the other two pre-publication texts.

Among the evidence of a link between the *Dear Boss* letter and the *Saucy Jacky* postcard,
the strongest piece of evidence is the presence of a shared distinctive 4-gram, *letter back till I*. The syntactic structure underlying this 4-gram is a verb phrase headed by a phrasal verb that, used within that particular structure underlying that particular unit of meaning, is also rare and distinctive overall. The presence of this 4-gram and of this structure thus supports the hypothesis that the two texts were written by the same person. This conclusion is substantiated by the fact that despite the presence of about two hundred texts trying to imitate the style of the *Dear Boss* letter or *Saucy Jacky* postcard, no other text has managed to reproduce this structure or 4-gram, which indeed this analysis has proved to be the real distinctive feature of these two texts.

The only exception is the *Moab and Midian* letter, which does not use the 4-gram but contains an instance of *keep back* meaning ‘to withhold’, including the co-selection of the position of the object and of the adverbial clause introduced by *till*. Furthermore, the *Moab and Midian* letter also shares another distinctive lexicogrammatical structure with *Saucy Jacky*, the verbless clause ADJ–ble event this time which elaborates the previous clause ending with the adverb *tomorrow*. It is not possible to discount that the author of this letter was simply more skilled in copying the style of *Dear Boss* than others, as by the time the *Moab and Midian* letter was received all the earliest texts were publicly available. However, the *Moab and Midian* letter is striking in also being the most similar letter in terms of the number of shared word 2-grams, even despite the fact that probably hundreds of other authors tried to imitate the style of *Dear Boss* and *Saucy Jacky*.

The analysis also points out that there is no link between the *From Hell* letter and the other historically important texts in the case. Although this lack of link does not constitute evidence that they were *not* written by the same person, this finding does lend some support to the initial presuppositions of other scholars that *Dear Boss* and *Saucy Jacky* are independent from the *From Hell* letter (Rumbelow, 1979). This and many other letters in the JRC texts can be analysed in more detail in the future.

Historically speaking, the comparison presented between the earliest letters ever received in the Whitechapel murders case provide linguistic evidence supporting the hypothesis that the two
most iconic texts sent during the case were written by the same person. Although several scholars have already commented on the similarity of the handwriting of the Dear Boss letter and the Saucy Jacky postcard, the common authorship of these two texts has never been established with certainty. The present analysis, however, found linguistic evidence that supports the common authorship of these two texts. Future analyses focused on their profiling or on the comparison with known writings of suspect authors can thus take as point of departure a link between these two texts.

Additionally, of great historical importance is also the link found between the two earlier iconic texts and the Moab and Midian letter, since this text is one of the most controversial in the JRC. Beside being the third and last letter that was ever sent to the Central News Agency, after Dear Boss and Saucy Jacky, Bulling's decision of sending a copy of the Moab and Midian letter instead of the original was never justified by the journalist and still remains suspiciously unexplained (Evans and Skinner, 2001). The linguistic link found between these three texts is therefore far from coincidental in the light of the other non-linguistic evidence and significantly contributes to the debate on the origin of the letter.

The present analysis is also successful in presenting serious implications for modern research in forensic linguistics and authorship analysis. The JRC is a corpus made up of texts the majority of which was fabricated by individuals that were imitating the style of the Dear Boss letter and of the Saucy Jacky postcard. However, it is evident that none of the authors of these texts successfully managed to individuate that the real linguistic distinctiveness consisted in a seemingly common string such as letter back till I, or in the phrasal verb keep back and its underlying structure, or even in simply the presence of the meaning of ‘withhold this letter’, found in only two other Jack the Ripper texts but encoded differently.

Instead, impostors imitated structures such as the salutation Dear Boss. Quantitatively speaking, despite the presence of these letters in full in the public domain, only a very limited percentage of them presents substantial linguistic similarities, implying that techniques such as the analysis of short texts using similarity measures such as the Jaccard coefficient are quite effective in
filtering this type of noise.

Theoretically, the results presented in this paper also contribute to the understanding of idiolect. A superficial reading of most of the JRC letters would only reveal their similarities in terms of meanings, themes, purposes, and some phraseology. However, this analysis has revealed that by investigating the way these meanings, themes, and purposes are encoded linguistically uniqueness emerges, as demonstrated by the relatively low average Jaccard distances between the letters. As shown by Wright (2017) for short emails, although meanings and speech acts can be shared, it is the way they are encoded in words and syntactic structures that tends to be idiosyncratic or unique.

1.6. Conclusions

In this paper, an analysis of the texts sent during the Whitechapel murders case was presented. This analysis found linguistic evidence that supports the hypothesis that the two most iconic texts signed as ‘Jack the Ripper’, the Dear Boss letter and the Saucy Jacky postcard, have been written by the same person. Because of the number and the distinctiveness of the linguistic similarities, it is likely that an authorial link also exists between these two texts and a third letter sent to the same recipient, the Moab and Midian letter. These results constitute new forensic evidence in the Jack the Ripper case after more than one hundred years, even though they do not reveal information about the identity of the killer(s).

Besides the historical and forensic implications, the results presented in this paper also have interesting consequences for modern research in authorship analysis, forensic linguistics, and research on idiolect. The results in this paper present additional evidence that uniqueness in linguistic production can be found in the way meaning is encoded and that this encoding of meaning can be difficult to imitate.
1.7. References


