Language and Law / Linguagem e Direito, *Vol. 11*(2), 2024, pp. 160-178
DOI: doi.org/10.21747/21833745/lanlaw/11_2a7

"This is an extortion note": a corpus-driven genre analysis of commercial extortion letters†

Marton Petyko[1], Lucia Busso[2], Sarah Atkins[3], Emily Chiang[4], Nabanita Basu[5], and Tim Grant[6]

[1]*Aston Institute for Forensic Linguistics, Aston University*
[2]*Aston Institute for Forensic Linguistics, Aston University*, l.busso@aston.ac.uk
[3]*Aston Institute for Forensic Linguistics, Aston University*, s.atkins@aston.ac.uk
[4]*Aston Institute for Forensic Linguistics, Aston University*, e.chiang2@aston.ac.uk
[5]*Northumbria University*, nabanita.basu@northumbria.ac.uk
[6]*Aston Institute for Forensic Linguistics, Aston University*, t.d.grant@aston.ac.uk

Abstract
This study presents a corpus-assisted genre analysis of 39 British commercial extortion letters, the largest dataset of its kind. We use a combined methodology that triangulates evidence from move analysis, corpus linguistics, and clustering algorithms. We innovate on traditional move analysis by using clauses as the unit of analysis and interrater reliability testing to identify 11 rhetorical moves. Corpus linguistic methods, including n-gram and collocation analysis, reveal that while certain moves (Threat and Demand) are highly prevalent and "core" for the genre of extortion letter, there is no consistent structural pattern across texts. Instead, rhetorical moves occur flexibly in texts, with no recurrent patterns, reflecting their status as an "illicit genre" produced outside conventional discourse communities. We further use clustering algorithms to explore whether we can reliably differentiate letters based on moves frequency. This multi-method approach demonstrates that rhetorical move prevalence, rather than sequence, can serve as a robust basis for genre classification in extortion communications. Furthermore, the study offers methodological innovations for forensic linguistics and genre.
Keywords: *extortion letters; malicious communications; move analysis; corpus linguistics; hierarchical clustering.*

†The present paper is the last scholarly work of Dr. Marton Petyko, who passed away in November 2023 and is greatly missed by his colleagues and friends at the Aston Institute for Forensic Linguistics. Paragraph 4, 5, and 5.1 are still largely his own writing, and he carried out the corpus linguistic analysis here presented. Lucia Busso is responsible for the move maps and paragraph 2 and 5.2, and the writing of paragraph 5.3. The analysis in 5.3 is done by Nabanita Basu. The Move Analysis, the coding, and the writing of the rest of the paper is the result of a collaboration among the first 5 authors.

Resumo
Este estudo apresenta uma análise de género textual baseada num corpus de 39 cartas de extorsão comercial britânicas, o maior conjunto de dados deste género. Utilizamos uma metodologia combinada que reúne e relaciona dados de análise de movimentos, linguística de corpus e algoritmos de agrupamento ("clustering"). Destacamo-nos da análise tradicional de movimentos, uma vez que utilizamos orações como unidade de análise e testes de fiabilidade interavaliadores para identificar 11 movimentos retóricos. Os métodos linguísticos do corpus, incluindo a análise de n-gramas e de colocações, revelam que, embora certos movimentos (Ameaça e Exigência) sejam altamente prevalecentes e “nucleares” para o género da carta de extorsão, não existe um padrão estrutural consistente nos textos. Em vez disso, os movimentos retóricos ocorrem de forma flexível nos textos, sem padrões recorrentes, refletindo o seu estatuto de “género ilícito” produzido fora das comunidades discursivas convencionais. Além disso, utilizamos algoritmos de agrupamento para explorar a possibilidade de diferenciar as cartas de forma fiável com base na frequência dos movimentos. Esta abordagem multimétodo demonstra que a prevalência de movimentos retóricos, e não a sequência, pode servir como uma base sólida para a classificação de géneros em comunicações de extorsão. Além disso, o estudo oferece inovações metodológicas para a linguística forense e para o estudo dos géneros textuais.
Palavras-chave: *Cartas de extorsão; Comunicações maliciosas; Análise de movimentos; Linguística de corpus; Agrupamento hierárquico.*

1. Introduction

Malicious and threatening communications, of which written extortion letters are a particular type, have been a key area of study for forensic linguistics. Scholars have analysed different types of malicious communication in different ways, using SFL and Appraisal (Hurt, 2020; Reczek, 2023), Biber’s Multidimensional Analysis (Nini, 2017, 2019), Speech Act theory (Spitzberg & Gawron, 2016) and Move Analysis (Abaalkhail, 2015; Chiang & Grant, 2019). Threatening language characteristics have also been analysed using a plethora of different data types, from Late Modern English threatening letters (Neumaier, 2025) to spoken threats (Tompkinson, 2023), to a large corpus of different types of threats (Gales, 2015).

However, these analyses have not focused on the genre of extortion letters specifically, and the question of whether this text type constitutes a genre has remained relatively underexplored. Some studies have identified that the form or ‘text type’ of extortion communications is only minimally shaped by norms, with writers borrowing from other genres, such as business letters, but with the texts open to a large degree of individual variation (e.g., Fobbe, 2011, 2020). However, such studies are often necessarily limited to small numbers of texts, mainly because of the difficulty of accessing texts beyond single case studies, and there is no study – to the best of the authors' knowledge – that examines more than a single digit number of texts.

The current study is therefore the first of its kind, as we analyse the largest existing collection of 39 written commercial extortion communications (letters and emails), as all other works (to the best of the authors' knowledge) analyse single digit amounts of documents. The documents all come from a UK context, and span from 2008 to 2019. With this in mind, we address the following research questions:

1. What are the recurring communicative functional units in commercial extortion communications?
2. Can we describe the discourse structure of this genre according to these units?
3. How homogeneous/varied is the collection of texts?

In the remainder of the paper we first discuss the two key theoretical concepts for our analysis: genre (section 2) and rhetorical moves (section 3). We then introduce the data (section 4) and present the methodology and results from our Move Analysis in section 5, looking at move prevalence (5.1), patterns (5.2), and clustering (5.3).

2. Genre analysis and forensic texts

The concept of genre is widely debated in the academic literature, with three major research traditions shaping genre studies (Xia, 2020). The English for Specific Purposes (ESP) approach, notably Swales (1990), views genre as a class of communicative events with shared purposes, shaped by discourse communities. Systemic Functional Linguistics (SFL, Martin, Christie, Rothery, & Reid, 1987), sees genre as a staged, goal-oriented social process. Finally, Rhetorical Genre Studies (RGS), following Miller (1984), frames genre as a response to recurring rhetorical situations, focusing on the typification of rhetorical actions.

Despite their differences – ESP's focus on community, SFL's on social processes, and RGS on rhetorical recurrence – there is broad consensus that genre is inherently social and contextual. Understanding genre, across these frameworks, involves analysing how texts reflect the practices and goals of the communities that produce and use them (Bhatia, 2016; Corbett, 2006). Biber and Conrad (2009) emphasise this social dimension, defining genre as a conventionalised construct, characterised by consistent linguistic features and a unified communicative purpose.

This study adopts Swales (1990) ESP-based definition, particularly useful for analysing recurring rhetorical moves in extortion communications. Swales defines genres as classes "of communicative events, the members of which share some set of communicative purposes", recognised by expert members of the parent discourse community (Swales 1990, p. 58). Central to his model are the notions of 'discourse community' ('socio-rhetorical networks that form to work towards sets of common goals', Swales, 1990, p. 9) and 'communicative purpose', which drives the language activities of the discourse community.

However, the notion of 'discourse' communities outlined above, potentially becomes problematic with forensic linguistic genres. Malicious communications, such as extortion letters, are unlikely to stem from a 'socio-rhetorical network' of authors (Swales, 1990, p. 9) nor from repeated exposure to examples within a discourse community. Bojsen-Møller, Auken, Devitt, and Christensen (2020) address this by proposing a fifth category - 'illicit genres' - to Miller's 2017 four genre typology. Illicit genres are not learned institutionally, lack formal instruction, and are socially marginal (Bojsen-Møller et al., 2020, pp. 34–35). Consequently, these genres are recognised as such not by a clear community of authors and users, but rather by "uptake communities", that is audiences, such as law enforcement or victims - those who label and respond to these texts, whether or not they are the intended audience. 'Uptakes' can be expected or unexpected, welcome or unwelcome (Freadman, 2012), which, in the case of illicit genres, Bojsen-Møller et al. 2020 suggest, results in a 'dual addressivity', simultaneously targeting victims as recipients while anticipating interception by authorities. This notion of 'uptake community' underpins our analysis of genre, where we focus on functions and actions accomplished within the texts.

3. Move analysis

Move analysis is a functional framework for examining the conventional discourse structures within genres. 'Moves' are rhetorical or discoursal units that serve coherent communicative functions, often realized through combinations of lower-level strategies (Biber, Connor, & Upton, 2007; Swales, 2004). Originally developed for academic writing, move analysis has since been applied to a range of genres, e.g; marketing texts (e.g., Campbell & Naidoo, 2017), online reviews (Skalicky, 2013) and crowdfunding texts (Liu & Deng, 2016). Despite this utility, theoretical and methodological challenges persist, particularly in adequately defining 'communicative function' for identifying moves, with similar terms such as rhetorical function, communicative purpose and communicative intention, are often used interchangeably (Askehave & Swales, 2001; Bhatia, 1996). While individual moves are often richly described, methods for identifying them remain opaque, limiting replicability. Moreno and Swales (2018) propose a bottom-up approach, identifying steps and sub-moves first and only later categorizing these into higher-level moves. However, their functional approach lacks consistent, formal criteria.

Another challenge lies in using moves to define genre membership. Early studies often relied on the presence of 'obligatory' and 'optional' moves, particularly in academic genres (Biber et al., 2007; Swales, 2004). However, in more fluid genres, researchers have tended towards more flexible terms, for example Samraj and Gawron (2015, p. 95), found that no move occurs in 100% of suicide notes. Instead, they propose identifying 'core' moves and 'minor' moves based on frequency. Building on Samraj and Gawron's concept, Chiang (2018) extends this to online child abuse conversations, using move frequency to assess "coreness" and their relevance to the text type. These studies suggest that 'occluded' genres, lacking fixed moves or structure, require a more flexible model of move analysis .

Our study refines move analysis for forensic linguistic application in two ways. Firstly, by using clauses as the unit of analysis and introducing inter-rater agreement as a novel method to identify and validate rhetorical moves. Secondly, we follow Chiang (2018) in using a three-tiered classification system for moves – 'core' (nearly 100% of texts), 'typical' (at least 50% of texts) and 'atypical' (under 50% of texts). This approach enhances the reliability of analysis for 'occluded' forensic genres.

4. Data and methods

The 39 written extortion communications were provided by a UK law enforcement partner. All had been previously classified and investigated as 'extortion communications' by the police, indicating an uptake community that had recognised and labelled these texts as belonging to the same genre. The texts all relate to cases that the law enforcement partner classed as 'historic' (at least 12-months old) and either resolved ('historic–complete'), or unresolved but no longer under investigation ('historic–incomplete').

The study examines the rhetorical structure of the letters, without inferring authorial or case-specific details, due to limited metadata (e.g. on number of perpetrators or extortion outcomes). The only reliable metadata was the police subdivision into individual communications and series (i.e. multiple letters by same author).

Ethical considerations were central, particularly in protecting the confidentiality of victims and organizations mentioned for the 'historic–incomplete' cases that had not entered the public domain. Ethical approval was obtained from Aston University as well as compliance with the law enforcement partners' protocols. Anonymisation was key to preventing traceability, involving secure storage and systematic redaction of identifiers (e.g. replacing names with <MaleFirstName1>, <Town>, etc). This process, developed in consultation with the police, followed best practices in anonymising sensitive linguistic data. The final dataset comprises 39 texts totalling 9,295 words, with texts ranging from 31 to 951 words (average length: 270 words, standard deviation: 222.2). Figure 1 plots text length range in the corpus.

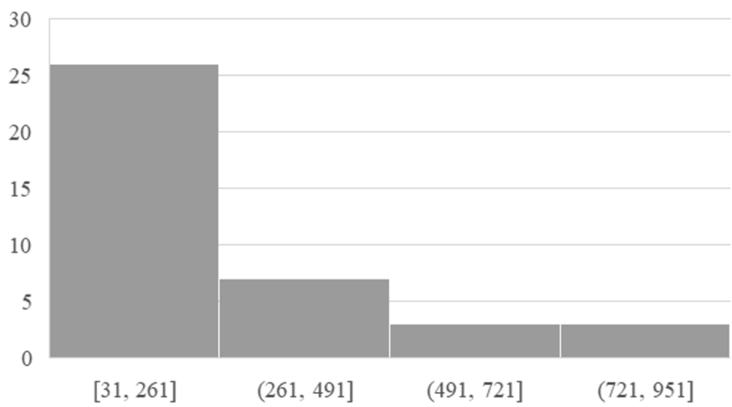

Figure 1. Text length in the corpus. As it can be seen, texts are generally short, between 31 and 261 words long

The dataset was analysed combining qualitative, corpus-based, and computational methods. Specifically, the letters were coded for rhetorical moves using a new and more reliable system that uses clauses as the basic unit of analysis and that involve iterative coding and reliability testing. We first selected five texts randomly and broke them down to clause units. Then four members of the research group independently analysed these extortion communications and created preliminary lists of recurring moves. We compared these preliminary lists and developed a refined set of 11 key moves. Once the move set had been finalised, we wrote a codebook with definitions and examples for every move identified (Section 5). Four members of the research team then coded moves in the five selected texts and we then calculated the level of interrater reliability for this coding process (section 5.1).

Section 5.2 introduces a corpus-based analysis of move n-grams and collocations to explore how moves pattern in the letters. Section 5.3 uses clustering algorithms to group letters in the corpus in statistically stable and coherent groups based on move prevalence.

5. The rhetorical moves in commercial extortion communications

Using an iterative annotation method, four coders from the research group independently coded five randomly selected texts, broken down into clauses (12% of the corpus, 20% of the total clauses). Inconsistencies were discussed and a refined, final set of 11 key moves was agreed. This move codebook includes *Opening, Sign-Off, Demand, Instructions, Justification, Threat, Demonstrating Credibility, Consequences, Additional Persuasion, Statement of Purpose, and Pre-Announcement.*

Openings and **Sign-Offs**, similar to other types of written communications, mark the beginning and the end of the analysed text. Some display more conventional formality, such as *Dear <supermarket> directors* [S02L01C01][1] or *Good luck as I await your reply* [S22L01C49], while others are informal, such as *Hello*, [S17L01C01] and *Bye luvvie* [S18L01C27].

Demands are central to extortion letters, urging the addressee to make a payment, *You must pay £50,000 in cash on Sunday 14th May* [S08L01C13] and *I demand a sum of 250.000 £ in my bitcoin account by the end of December* [S02L01C02] or respond to the author (*Get back to me now if you are ready to pay some fees to spare your life* [S22L01C31] and *Do not ignore this threat* [S13L01C08]). Authors often instruct the addressee not to contact the authorities (*Do not go to the police* [S23L01C12]).

Instructions frequently accompany demands, guiding the addressee on how to deliver payment or perform additional actions. These instructions can be highly detailed and sometimes make up a large part of the text, as the step-by-step instruction clauses in this text illustrate:

> *At 1400hrs you will drive to the <Supermarket1> car park and stop at the unloading only at the front of the store. Wait for 5 mins then head for <StreetName1>. You will then drive up to <HouseNumber1> <StreetName1> until you see an off licence on your left. Take an immediate left turn at this off licence and park up against the shop wall. You will see a brick construction halfway across the wall 4 bricks high and rectangle in shape. It surrounds a steel grate. There are bricks inside this. Lift up the corner brick and place package there. Get in your car and drive off. [S15L01C50-60]*

Justifications provide reasons for the demand, acting as a persuasive technique on the authors' part to push the addressees towards compliance. For example, extortionists may claim that their victim owes them money (*I feel a little disappointed that you do not seem to be willing to pay me back my £10000 you owe me* [S10L01C03]), previously caused harm (*you kindly made me unemployed* [S17L01C06]) or committed a

[1] The shorthands between square brackets refer to the location of the examples in our corpus. For example, S02L01C01 means that the example comes from the first clause of the first letter of the second letter series. (S stands for series, L for letter and C for clause.)

crime (*Several weeks ago you purchased items of jewellery that were stolen during a robbery at a couple's home. Five days later the lady suffered a stroke and to all intents and purposes their lives have been devastated. I must state that I don't hold you responsible for these events only to the facts of you buying stolen property* [S15L01C1-3]), thus framing demands (and their subsequent threats) as legitimate and morally justifiable.

Threats are the other fundamental move of extortion communications, used to coerce compliance. Threats were classed as a statement that expresses the extortionist's willingness or intent to cause physical harm (*Get back to me now if you are ready to pay some fees to spare your life. If you are not ready for my help, then I will carry on with my job straight up* [S19L01C32-34]; *You will not be harmed, but your wife, Sons, business partners and employees will be* [S08L01C7-8]) or damage the business interests of the targeted company (*On the 5th December 2013 your <supermarket> products contaminated with salmonella will be simultaneously placed in 25 major retail outlets around the country* [S13L01C01]).

Demonstrating Credibility was performed to support threats, with extortionists claiming access to the necessary tools (*We have devices ready to cease your production for one day – one week – one month – three months and more* [S11L01C12]), knowledge (*As I am writing to you now my men are monitoring you and they are telling me everything about you* [S22L01C27-28]), expertise (*we would detect [the tracking device] before we got close* [S09L01C48]), and determination (*Secondly I am trained to follow through all plans I make and will not be deterred from anything or anyone in the in the pursuit of those plans* [S15L01C15-16]). These portray the authors as experienced criminals, often with claims to membership of organised crime groups.

Consequences contrast the costs of non-compliance (*However if you put us to this trouble we double what we want to 200,000 pounds* [S26L01C21-22]) with the benefits of compliance, such as promises that no further demand will be made (*When you deliver the money I will consider the matter closed. There will be no further contact between us* [S15L01C31-32]). This move also covers a 'cost/benefit analysis' of compliance, where authors tend to minimise the costs (*<company name> as a company is financially stable and should be able to pay* [S12L02C7-8]) and maximise the benefits of compliance (*you can quickly protect the public* [S26L01C17]).

Additional Persuasion, encompasses a variety of persuasive techniques, grouped into a single move for practical operation of the analysis. This included non-negotiability (*This is a serious requirement and is not negotiable* [S09L01C7-8]), appeal to reason (but stupid you are not [S05L01C24]), or to the futility of resistance (*You cannot win* [S10L02C69]), appeals to moral duty (*so the faith or your employees and customers is in your hands* [S02L01C05]), and follow-up promises (*We will be in touch* [S07L01C19]).

Statement of Purpose declare the communicative purpose of the texts (*Subject: extortion message* [S20L01C06]), suggesting some awareness of 'genre' by extortionists themselves.

Pre-Announcements introduce the move that will follow, such as the heading *Instructions* [S08L01C19] before the instructions.

5.1. Prevalence of moves

Once the codebook had been finalised, the sample of texts was then re-coded, achieving high interrater reliability using percentage agreement (88% of codes with 75% agreement), following Rau and Shih’s 2021, who argue this is more appropriate for this type of unit-boundary coding[2].

A trained, external coder was then introduced, who first coded the same five texts.[3] The high level of agreement was maintained, with 84% of clauses achieving agreement above 75% and 68.7% in 100% agreement between all coders. The coder then independently annotated the rest of the corpus, with ongoing discussions with the research team to ensure consistency.

Table 1 shows the range (percentage of texts where a move appears at least once) and frequency (percentage of clauses for each move across all texts) of the 11 rhetorical moves across the 39 texts, divided into 1,062 clauses.

Table 1. The range and frequency of moves in the corpus

Move	Range (n = 39)	Frequency (n = 1,062)
Threat	92.3%	15.7%
Demand	92.3%	9.3%
Consequence	87.2%	15.4%
Opening	87.2%	5.6%
Instruction	84.6%	14.8%
Demonstrating credibility	82.1%	19.4%
Additional persuasion	64.1%	7.5%
Sign-off	51.3%	2.5%
Pre-announcement	46.2%	2.8%
Justification	35.9%	6%
Statement of purpose	20.5%	0.8%

Results show threats and demands are the two core moves, found in 92.3% of texts. Demands appear in every stand-alone written text and multi-text series, while Threats also occur in every multi-text series and in almost all stand-alone communications, except one in which the author implicitly implies intimidation by referencing damage to an organised crime member’s car.

Beyond these two core moves, other typical moves (Consequences, Instructions, Demonstrating the Credibility, Openings, Additional Persuasion, and Sign-Offs) appear in more than three-quarters of the texts. The prevalence of discussing Consequences and Demonstrating Credibility, along with Additional Persuasion, suggests that while threats are a fundamental persuasive tool, extortion communications employ a broader range of persuasive techniques.

[2]Percentage agreement is calculated in a pairwise manner, i.e. between two annotators at a time. That is, the coding of Annotator 1 is compared to the coding of Annotator 2, then to the coding of Annotator 3, and so on. If the codings of the two annotators coincide, a value of 1 is assigned to the cell. If the codings diverge, it is assigned 0. This process is repeated for every annotator, for a total of 6 comparisons (1 vs 2, 1 vs 3, 1 vs 4, 2 vs 3, 3 vs 4). The total is then divided by the total number of comparisons and assigned a percentage score. A score of 100% is given to instances in which all 4 annotators agree, 75% is given to cases in which 3 out of 4 annotators agree, and so on.

[3]We follow the rule of thumb followed by the literature that values greater than 0.70 are typically acceptable for consistency estimates of interrater reliability (Stemler, 2004).

Less surprising is the widespread presence of Openings and Sign-Offs, reflecting norms in written communications more generally, while the prevalence of Additional Persuasion aligns with the generally persuasive nature of extortion communication.

Finally, three of the moves identified, Pre-Announcement, Justification, and Statement of Purpose, can be considered atypical, present in less than 50% of the texts. Justification therefore emerges as the least common persuasive technique.

The mosaic plot in Figure 2 visualises the frequency of each move within individual texts, showing that frequency of moves, especially the core and typical moves, varies greatly across individual texts. For example, Demonstrating Credibility (in orange) is rather dominant in texts 13, 25 and 39, but it becomes much less prevalent in texts 9, 32 and 36.

Figure 2. Prevalence of moves within individual texts. Thickness of bars refer to length of the text (longer texts have thicker bars). The moves are shown in the legend in alphabetical order: AdP - Additional Persuasion, Con - Consequences, Dem - Demand, DemCred - Demonstrating Credibility, In - Instructions, Jus - Justification, Op - Openings, PrA - Pre-announcement, SO - Sign offs, SoP - Statement of Purpose, Th - Threat.

5.2. Move patterns

Having coded moves across the corpus, we use corpus linguistics to explore patterns in the sequencing of moves. The importance of using corpus linguistic methods to enhance and complement discourse studies has been highlighted in the literature (Biber et al., 2007; Upton & Cohen, 2009), as it enables

> “detailed but generalizable analyses of discourse structure across a representative sample of texts from a genre. (....) It is then possible to identify the sequences of move types that are typical for a genre, and against that background it is also possible to identify particular texts that use more innovative sequences of move types.” (Upton & Cohen, 2009, p. 588).

Recurrent structures of moves have been found in many professional contexts, such as promotion letters and job application letters (Bhatia, 1993), and business letters of negotiation (Dos Santos, 2002). However, when the method is applied to forensic texts, linear structures in the moves are not always found. For example, Samraj and Gawron’s 2015 study of suicide notes found texts only required one move from a set of ‘core’ moves to be included in the ‘genre’ and that they displayed a great variation in structure.

Having access to a relatively large corpus of extortion letters from a variety of different authors gives us the perfect testbed to explore whether any recurrent patterns of moves can be found in this illicit genre. To do so, we employ corpus linguistics methods using LancsBox (Brezina, Weill-Tessier, & McEnery, 2020). For these analyses we consider moves as a whole, and not clauses, as units. That is, a move that spans multiple clauses is considered to be one unit (leaving us with 612 moves across 39 texts). So, example 1.(a)-(c) below (taken from text 008) consists of 3 consecutive clauses which were all coded as one move (Threat).

1. (a) *The consequences will be Very serious.*
   (b) *You will not be harmed,*
   (c) *but your wife, Sons, business partners and employees will be.*

We analyse patterns of moves using ngram and collocation analysis. We use ngram analysis to explore sequences of two or three adjacent moves, while collocation analysis is instead focused on the co-occurrence of two non-adjacent moves within a 2L 2R window span (i.e., 2 moves to the left 2 moves to the right). The ngram analysis of move sequences allowed us to find fixed patterns, whereas we used the collocation approach to explore looser associations between two moves, where two moves can stand directly next to each other or can have any third move between them. The minimum raw frequency in both ngrams and collocation analysis was 1. After applying the above criteria, we found 392 move sequences and 142 collocations.

Table 2. The ten most frequent move ngrams in the corpus

Move sequence	**Frequency**	**Range** (n = 39)
Consequence -Threat	22	43.6%
Demonstrating Credibility -Threat	22	30.8%
Consequence - Additional persuasion	20	33.3%
Threat - Consequence	20	35.9%
Demand - Instruction	19	38.5%
Demand - Consequence	19	33.3%
Demonstrating - Credibility Consequence	18	33.3%
Consequence - Demonstrating credibility	17	33.3%
Instruction - Demand	16	33.3%
Instruction - Consequence	15	30.8%

Table 2 shows the ten most frequent move sequences in the corpus, all of which are bigrams (i.e. sequences of two moves). As expected, the most frequent sequences are made up of the most frequent moves. For example, the joint most frequent sequence, *Consequence-Threat* and *Demonstrating Credibility-Threat*, feature the three most frequent moves: *Demonstrating Credibility* (19.4%), *Threat* (15.7%), and *Consequence* (15.4%). This suggests that there are no preferred sequential patterns of moves in our corpus, and that on the contrary moves occur quite freely in texts.

Moreover, none of the sequences can be considered typical, as even the most widespread one, *Consequence-Threat*, only appears in 43.6% of the texts, while the rest occur in around only one third of the extortion notes. The ngram analysis also suggests that the moves do not follow any specific order in the analysed texts. For example, *Consequence* is followed by *Threat* 22 times while *Threat* is followed by *Consequence* 20 times. Similarly, *Instruction* comes after *Demand* 19 times and *Demand* after *Instruction* 16 times.

These results overall indicate that while the analysed extortion communications do have core and typical moves, these moves do not form recurrent structural patterns. In other words, while genres with a strong community of practice generally show a fixed order of recurrent move sequences (e.g. academic articles, Swales 1990), the generic structure of this text type appears to be driven by function rather than by a fixed order of recurrent move sequences. In this sense, the common features of the genres are more functional than formal. This raises interesting questions concerning the notion of genre based on observation of the same functions across texts, rather than the ordering or arrangement of these functions. This finding is in line with other works on genres that do not have a clear community of practice, such as the already mentioned work by Samraj and Gawron (2015) on suicide notes and this clearly deserves further investigations.

Table 3. The ten most frequent move collocations in the corpus

Move collocation	Frequency
Consequence […] Threat	66
Consequence […] Demonstrating credibility	55
Demand […] Consequence	53
Demonstrating credibility […] Threat	49
Demand […] Instruction	46
Consequence […] Additional persuasion	45
Threat […] Demand	43
Instruction […] Consequence	38
Threat […] Instruction	38
Instruction […] Demonstrating Credibility	36

Table 3 reports the ten most frequent move collocations in the analysed texts, which reinforce the key findings of the ngram analysis. While the most frequent moves form the top collocations, the collocations themselves are not particularly frequent, albeit more common than the sequences, suggesting that the moves do not form typical patterns in the texts.

Figures 3 and 4 show move maps of the corpus. Move maps are colour-coded visual representations of the move structure and patterns in each text of the datasets (Chiang & Grant, 2017, 2019). In Figure 3, each move is represented by a different colour, and horizontal bars represent individual texts. It is immediately clear by looking at the colour distribution that very little stable patterning is present, apart from Openings and Sign-Off. On the contrary, there is great variety in the possible order that the moves can take across different texts. Figure 4 is a different take on a move map, showing the positions each individual move can take in the texts. It importantly reveals that one of the main reasons why we were unable to find any typical move patterns is that most moves do not have any typical position in the texts in the first place.

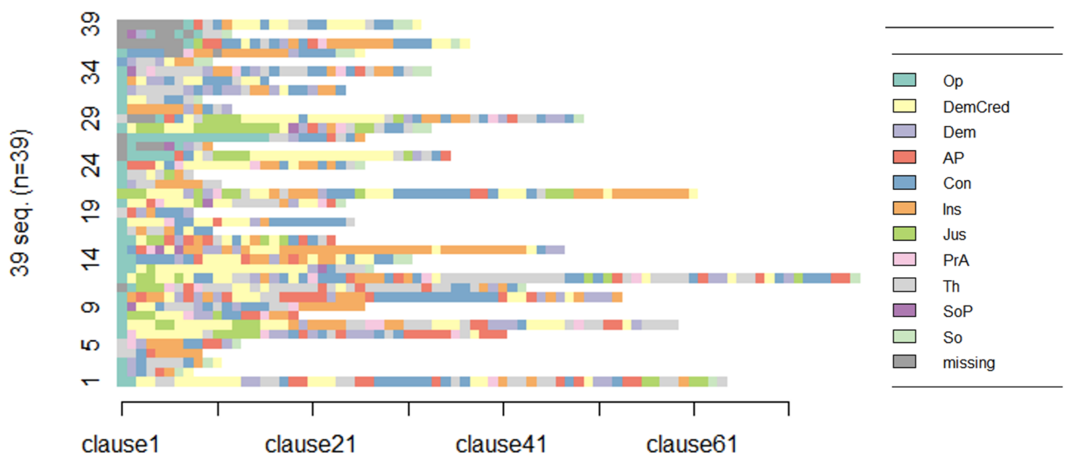

Figure 3. Move maps of commercial extortion communications. The moves in the legend are shown in alphabetical order: AdP - Additional Persuasion, Con - Consequences, Dem - Demand, DemCred - Demonstrating Credibility, In - Instructions, Jus - Justification, Op - Openings, PrA - Pre-announcement, SO - Sign offs, SoP - Statement of Purpose, Th - Threat.

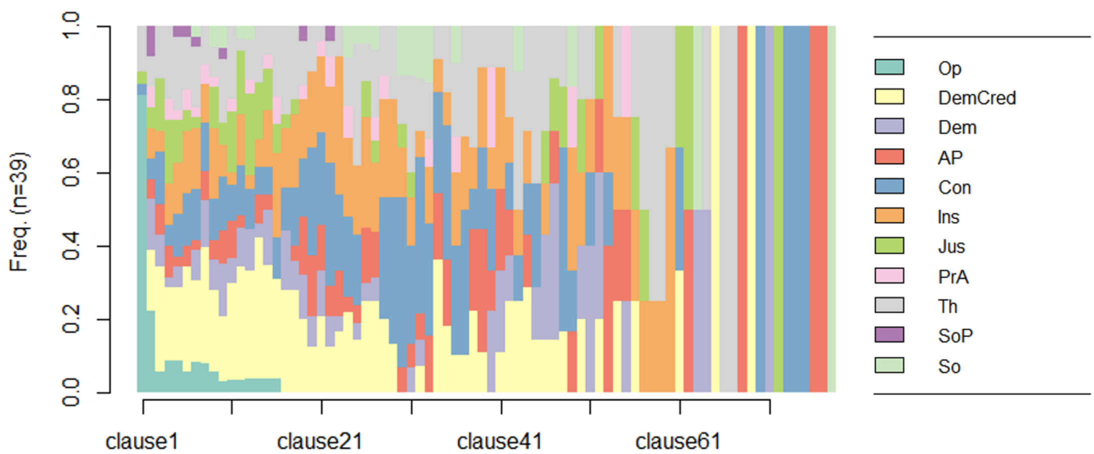

Figure 4. Positions of individual moves across communications. The moves in the legend are shown in alphabetical order: AdP - Additional Persuasion, Con - Consequences, Dem - Demand, DemCred - Demonstrating Credibility, In - Instructions, Jus - Justification, Op - Openings, PrA - Pre-announcement, SO - Sign offs, SoP - Statement of Purpose, Th - Threat.

Apart from Openings and Sign-Offs, which of course appear at the start and end of the communications, all the other moves can take any position. Threats for example (in gray at the top of the graph) can appear in virtually any position, from clause 1 to clause 70.

Overall, then, although the texts in this collection of extortion notes hold a similar communicative purpose and two core moves decisively emerge from the data, there is no consistent sequence or pattern to those moves. The relationship between the texts is perhaps of a somewhat 'looser kind' (Swales, 1990, p. 49) but they share a family resemblance through the core moves contained (Swales, 1990, p. 52), perhaps partly accounted for by the lack of a coherent discourse community within which these texts are written.

5.3. Clusters of moves

Since no dominant move pattern was found in the data, we turned to explore whether groups of texts can be identified within the corpus based on move frequency (and not positioning). The use of computational techniques that use rhetorical moves as input data is an innovative method first introduced by Chiang and Grant (2019). Here, we use clustering algorithms based on move frequency to explore whether texts can be grouped into statistically relevant and meaningful clusters based on move presence, absence, or frequency.[4] In other words, we want to see if moves can be reliably used not only as a descriptive tool, but also as a reliable variable for text differentiation. If that is the case, our more principled approach to move analysis can be potentially used as an investigative tool (similarly to what Chiang and Grant, 2019 propose).

As a preparatory step, we adopt term frequency-inverse document frequency (TF-IDF), which determines how relevant a word is to a particular document in view of the corpus under consideration. This is done by multiplying the frequency of the word in a document by the inverse document frequency of the word across the entire set of documents. In our case, instead of word frequency we use move frequency. Each text is hence defined by a set of 11 features, where each feature is the product of move frequency within each text, and the inverse frequency of texts in which each move appears. To prevent small feature values from being overshadowed by large ones, all features were scaled to mean value of zero and standard deviation of one, as a means of normalisation. We also estimated products of features by taking any 2, 3, 4, 5, 6, 7, 8, 9, 10 and 11 features at a time (Naseer, Noori, Qureshi, & Hong, 2016). Empirical selection was performed with a view to identifying explainable, logical patterns in the dataset. This procedure revealed that a combination of 6 and 7 features was optimal, hence we augmented our dataset with a combination of these feature products taking any 6 and 7 features at a time. Hence, the dataset consists of 39 texts each defined by a total of 803 features FORMULA.

To ensure reliability, three clustering algorithms were applied: *hierarchical clustering* (Murtagh & Legendre, 2014), *k-means clustering* (Hartigan & Wong, 1979) and *partition around medoids clustering* (*PAM*, Kaufman & Rousseeuw, 1990). The number of clusters tested ranged from 2 to 6, reflecting the corpus's limited size. Performance was evaluated using both interpretability and statistical metrics – connectedness, compactness, separation, and stability – using the R packages 'clValid' and 'fpc'(Brock, Pihur, Datta, & Datta, 2008; Hennig, 2020). These metrics assess (respectively) how well texts

[4] R code is available as supplemental material.

are grouped, how distinct the clusters are, and how consistent results remain under resampling. Clustering was based on Euclidean distance and Ward’s linkage, a method favoured in linguistic studies for producing balanced clusters.

Agglomerative hierarchical clustering with two clusters emerged as the most effective method, producing robust and interpretable groupings. Notably, both hierarchical and k-means clustering yielded identical results: one cluster contained two specific texts (S14L20 and S21L28), while the other included the remaining 37 (see Figure 4). This convergence, despite differing methodologies, suggests a strong underlying structure in the data (Handl, Knowles, & Kell, 2005).

The resulting clusters also demonstrated high stability scores (0.96 and 0.84), reinforcing the reliability of the findings and the presence of meaningful patterns in the corpus. Results show that cluster 1 consists of texts displaying all 11 moves, while texts in cluster 2 have at least one of the moves missing. This means that the cluster algorithm was able to group the texts based on the presence or absence of atypical moves. When we explored which of the atypical moves provide greater separation between the two clusters, we find that combinations including ‘Justification’ and ‘Statement of purpose’ make the greatest contribution. That is to say that the top 10 feature combinations which include those two moves increased the separation between the clusters.

This result suggests that the moves identified in the corpus have at least some predictive power. The difference between core and atypical moves seems to be crucial in separating groups of letters, which – in datasets with more reliable metadata – can be useful for investigative contexts.

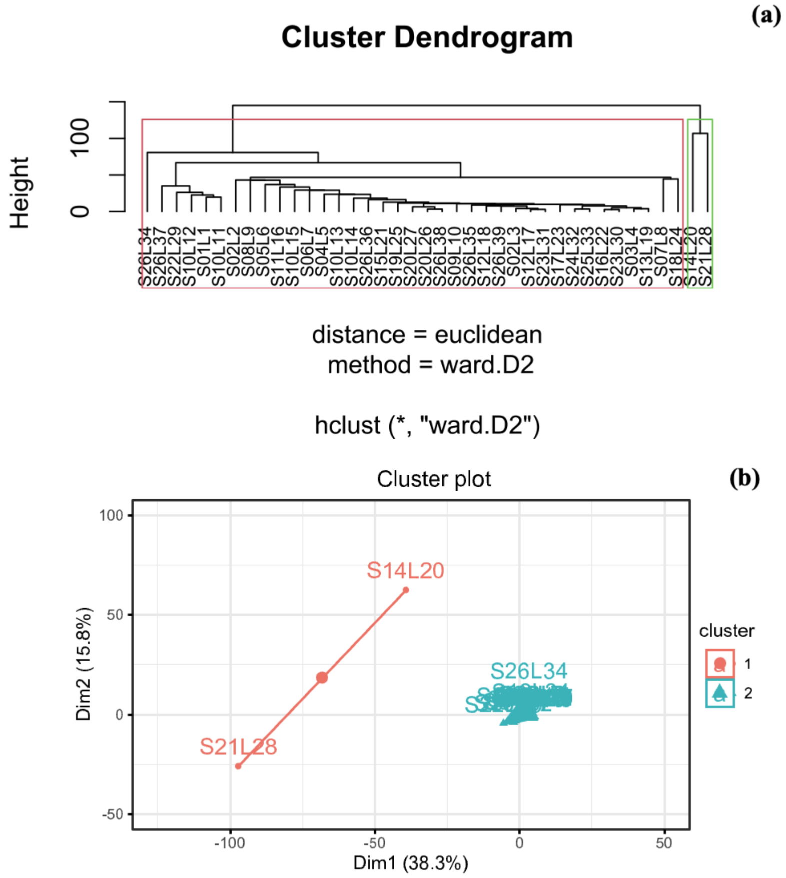

Figure 5. (a) Cluster dendogram for hierarchical agglomerative clustering. The labels represent identifier for each text in the corpus. The red and green rectangles represent the two clusters identified by the algorithm. (b) Scatter plot of the clusters identified by K-means (K=2) algorithm. Cluster 1 is represented in red and cluster 2 is represented in blue. The tightly overlapping labels for cluster 2 indicate that these data points all occupy the same area in the two-dimensional space.

6. Conclusions

This study provided an innovative approach to applying Move Analysis to the illicit genre (Bojsen-Møller et al., 2020) of commercial extortion letters, providing a noteworthy profile of the functional features in this genre. Specifically, we analyse the biggest corpus so far of commercial extortion letters, EXCrOW, with an improved version of Move Analysis. We tackle the lack of methodological clarity of the method by introducing clauses as the minimum unit of analysis and performing iterative coding and interrater reliability testing.

We found 11 rhetorical moves in our data: Opening, Sign-Off, Demand, Instructions, Justification, Threat, Demonstrating Credibility, Consequences, Additional Persuasion, Statement of Purpose, and Pre-Announcement. The two most prevalent moves identified are Threats and Demands, which appear in 92.3% of the texts. Beyond these, however, we have also found six typical moves which appear in more than half of the texts: Consequences, Openings, Instructions, Demonstrating Credibility, Additional Persuasion, and Sign-Off. Our analysis therefore highlights that while demands and threats are, perhaps inevitably, the most fundamental persuasive tools of the genre, authors also resort to a plethora of other communicative techniques, such as discussing Consequences of compliance and non-compliance, Demonstrating Credibility and Additional Persuasion. This suggests that in assessing the level of threat posed by an extortion letter, a more nuanced approach should be taken, considering not only explicit or implicit threats alone.

With regards to the discourse structure of commercial extortion communications, corpus linguistic analyses and move maps show that there are no discernible structural patterns in the data. Corpus-based analyses of moves – as discussed in Upton and Cohen (2009) – are essential for a more generalisable but at the same time fine-grained study of discourse structures and function. Extortion communications, like other illicit genres that lack a community of practice, seem to be better described in terms of prevalence rather than recurrent patterns of moves. In other words, while traditional genres like business letters or academic articles generally display a fixed structure, with identifiable structural patterns of moves, extortion communications exhibit a flexible internal structure. This is confirmed by our computational analysis, that using clustering algorithms demonstrates that meaningful distinctive groups of texts can be identified based on the presence or absence of rhetorical moves. We interpret these result as indicating that extortion as a genre is mainly driven by function rather than by strict internal formal constraints. This functional nature of extortion letters is due – we claim – to the absence of a coherent discourse community.

Our findings have important implications for the study of forensic genres at large. In fact, corpus-based analyses of recurrent patterns of discourse units (Biber et al., 2007; Upton and Cohen, 2009) work remarkably well for quantifying discourse information for most genres, but forensic genres – such as suicide notes or extortion letters – escape the notion of genre as made of recurrent and predictable patterns of discourse units. Rather, the rhetorical moves in these text types combine freely with no discernible pattern. Therefore, we argue that using clustering based on move prevalence – rather than methods rigidly based on sequencing – is an alternative quantitative analysis tool for illicit genres.

In terms of methodological implications, this paper intends to inform the methods of future studies that employ rhetorical moves. Firstly, using a consistent and easy-to-define unit of analysis (such as clauses) provides for a higher and better reliability to move analysis, making the method more principled and replicable. Secondly, the use of multiple coders and interrater reliability testing ensures that the coding of rhetorical moves remains principled and potentially replicable. Finally, we also combine qualitative move analysis with quantitative methods such as move maps, ngrams and collocation analysis of moves, and cluster analysis. This demonstrates that a principled move analysis can be successfully combined with quantitative tools to explore move patterns and positions within individual texts as well as distinctive groups of texts in a corpus.

References

Abaalkhail, A. (2015). *Rhetorical moves in an occluded genre: A qualitative analysis of suicide notes* (PhD dissertation). Carleton University.

Askehave, I., & Swales, J. M. (2001). Genre identification and communicative purpose: a problem and a possible solution. *Applied Linguistics*, *22*(2), 195–212.

Retrieved 2025-08-12, from https://academic.oup.com/applij/article-lookup/doi/10.1093/applin/22.2.195 doi: 10.1093/applin/22.2.195

Bhatia, V. K. (1993). *Analysing Genre.* Longman.

Bhatia, V. K. (1996). Methodological Issues in Genre Analysis. *Hermes, Journal of Linguistics*, *16*, 39–59. Retrieved 2025-08-12, from https://tidsskrift.dk/her/article/view/25383 doi: 10.7146/hjlcb.v9i16.25383

Bhatia, V. K. (2016). *Critical genre analysis: Investigating interdiscursive performance in professional practice.* Routledge.

Biber, D., Connor, U., & Upton, T. A. (2007). *Discourse on the Move: Using corpus analysis to describe discourse structure* (Vol. 28). Amsterdam: John Benjamins Publishing Company. Retrieved 2025-08-12, from http://www.jbe-platform.com/content/books/9789027291912 doi: 10.1075/scl.28

Biber, D., & Conrad, S. (2009). *Register, genre, and style.* Cambridge University Press.

Bojsen-Møller, M., Auken, S., Devitt, A. J., & Christensen, T. K. (2020). Illicit Genres: The Case of Threatening Communications. *Sakprosa*, *12*(1), 1–53. Retrieved 2025-08-12, from https://journals.uio.no/sakprosa/article/view/7416 doi: 10.5617/sakprosa.7416

Brezina, V., Weill-Tessier, P., & McEnery, T. (2020). *#LancsBox.*

Brock, G., Pihur, V., Datta, S., & Datta, S. (2008). **clValid** : An *R* Package for Cluster Validation. *Journal of Statistical Software*, *25*(4). Retrieved 2025-08-12, from http://www.jstatsoft.org/v25/i04/ doi: 10.18637/jss.v025.i04

Campbell, K. S., & Naidoo, J. S. (2017). Rhetorical Move Structure in High-Tech Marketing White Papers. *Journal of Business and Technical Communication*, *31*(1), 94–118. Retrieved 2025-08-12, from https://journals.sagepub.com/doi/10.1177/1050651916667532 doi: 10.1177/1050651916667532

Chiang, E. (2018). *Rhetorical moves and identity performance in online child sexual abuse interactions* (Unpublished doctoral dissertation). Aston University.

Chiang, E., & Grant, T. (2017). Online grooming: Moves and strategies. *Language and Law / Linguagem e Direito*, *4*(1), 103–141.

Chiang, E., & Grant, T. (2019). Deceptive Identity Performance: Offender Moves and Multiple Identities in Online Child Abuse Conversations. *Applied Linguistics*, *40*(4), 675–698. Retrieved 2025-08-12, from https://academic.oup.com/applij/article/40/4/675/4952155 doi: 10.1093/applin/amy007

Corbett, J. (2006). Genre and genre analysis. In K. Brown (Ed.), *Concise encyclopedia of pragmatics* (pp. 286–292). Elsevier.

Dos Santos, V. (2002). Genre analysis of business letters of negotiation. *English for Specific Purposes*, *21*(2), 167–199. Retrieved 2025-08-12, from https://linkinghub.elsevier.com/retrieve/pii/S0889490600000284 doi: 10.1016/S0889-4906(00)00028-4

Fobbe, E. (2011). *Forensische Linguistik. Eine Einführung.* Narr.

Fobbe, E. (2020). Text-Linguistic Analysis in Forensic Authorship Attribution. *International Journal of Language & Law*, *9*, 93–114. Retrieved 2025-08-12, from https://www.languageandlaw.eu/jll/article/view/78 (Publisher: International Language and Law Association (ILLA)) doi: 10.14762/JLL.2020.093

Freadman, A. (2012). The Traps and Trappings of Genre Theory. *Applied Linguistics*, *33*(5), 544–563. Retrieved 2025-08-12, from https://academic.oup.com/applij/

article-lookup/doi/10.1093/applin/ams050 doi: 10.1093/applin/ams050

Gales, T. (2015). The stance of stalking: a corpus-based analysis of grammatical markers of stance in threatening communications. *Corpora*, *10*(2), 171–200. Retrieved 2025-08-14, from https://www.euppublishing.com/doi/10.3366/cor.2015.0073 doi: 10.3366/cor.2015.0073

Handl, J., Knowles, J., & Kell, D. B. (2005). Computational cluster validation in post-genomic data analysis. *Bioinformatics*, *21*(15), 3201–3212. Retrieved 2025-08-12, from https://academic.oup.com/bioinformatics/article/21/15/3201/195678 doi: 10.1093/bioinformatics/bti517

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Applied Statistics*, *28*(1), 100. Retrieved 2025-08-12, from https://www.jstor.org/stable/10.2307/2346830?origin=crossref doi: 10.2307/2346830

Hennig, C. (2020). *Minkowski distances and standardisation for clustering and classification of high dimensional data.* arXiv. Retrieved 2025-08-14, from https://arxiv.org/abs/1911.13272 (Version Number: 2) doi: 10.48550/ARXIV.1911.13272

Hurt, M. (2020). *Pledging to Harm: A linguistic analysis of violent intent in threatening language* (PhD dissertation). Aston University.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis* (1st ed.). Wiley. Retrieved 2025-08-12, from https://onlinelibrary.wiley.com/doi/book/10.1002/9780470316801 doi: 10.1002/9780470316801

Liu, J., & Deng, L. (2016). A genre analysis of web-based crowdfunding discourse. *Asian ESP Journal*, *12*(2), 171–202.

Martin, J. R., Christie, F., Rothery, J., & Reid, I. (1987). Social process in education: A reply to Sawyer and Watson and others. In *The place of genre in learning: Current debates* (pp. 58–82). Deakin University.

Miller, C. R. (1984). Genre as social action. *Quarterly Journal of Speech*, *70*(2), 151–167. Retrieved 2025-08-12, from http://www.tandfonline.com/doi/abs/10.1080/00335638409383686 doi: 10.1080/00335638409383686

Miller, C. R. (2017). “Where Do Genres Come From?”. In C. R. Miller & A. R. Kelly (Eds.), *Emerging Genres in New Media Environments* (pp. 1–34). Cham: Springer International Publishing. Retrieved 2025-08-12, from http://link.springer.com/10.1007/978-3-319-40295-6_1 doi: 10.1007/978-3-319-40295-6_1

Moreno, A. I., & Swales, J. M. (2018). Strengthening move analysis methodology towards bridging the function-form gap. *English for Specific Purposes*, *50*, 40–63. Retrieved 2025-08-12, from https://linkinghub.elsevier.com/retrieve/pii/S0889490617303332 doi: 10.1016/j.esp.2017.11.006

Murtagh, F., & Legendre, P. (2014). Ward’s Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward’s Criterion? *Journal of Classification*, *31*(3), 274–295. Retrieved 2025-08-12, from http://link.springer.com/10.1007/s00357-014-9161-z doi: 10.1007/s00357-014-9161-z

Naseer, N., Noori, F. M., Qureshi, N. K., & Hong, K.-S. (2016). Determining Optimal Feature-Combination for LDA Classification of Functional Near-Infrared Spectroscopy Signals in Brain-Computer Interface Application. *Frontiers in Human Neuroscience*, *10*. Retrieved 2025-08-12, from http://journal.frontiersin.org/Article/10.3389/fnhum.2016.00237/abstract doi: 10.3389/fnhum.2016.00237

Neumaier, T. (2025). “I have come to the conclusion that you must die”: Threats in Late

Modern English threatening letters. *Journal of Historical Pragmatics*, *26*(2), 262–287. Retrieved 2025-08-12, from http://www.jbe-platform.com/content/journals/10.1075/jhp.23016.neu doi: 10.1075/jhp.23016.neu

Nini, A. (2017). Register variation in malicious forensic texts. *The International Journal of Speech, Language and the Law*, *24*(1), 99–126. Retrieved 2025-08-14, from https://utppublishing.com/doi/10.1558/ijsll.30173 doi: 10.1558/ijsll.30173

Nini, A. (2019). The Multi-Dimensional Analysis Tagger. In T. Berber Sardinha & M. Veirano Pinto (Eds.), *Multi-Dimensional Analysis: Research Methods and Current Issues* (pp. 67–94). Bloomsbury Publishing PLC.

Rau, G., & Shih, Y.-S. (2021). Evaluation of Cohen's kappa and other measures of inter-rater agreement for genre analysis and other nominal data. *Journal of English for Academic Purposes*, *53*, 101026. Retrieved 2025-08-12, from https://linkinghub.elsevier.com/retrieve/pii/S1475158521000709 doi: 10.1016/j.jeap.2021.101026

Reczek, A. P. (2023). *Forensic Applications of Appraisal Theory and Genre to Threatening and Malicious Language* (PhD dissertation). Aston University.

Samraj, B., & Gawron, J. M. (2015). The suicide note as a genre: Implications for genre theory. *Journal of English for Academic Purposes*, *19*, 88–101. Retrieved 2025-08-12, from https://linkinghub.elsevier.com/retrieve/pii/S1475158515000454 doi: 10.1016/j.jeap.2015.04.006

Skalicky, S. (2013). Was this analysis helpful? A genre analysis of the Amazon.com discourse community and its "most helpful" product reviews. *Discourse, Context & Media*, *2*(2), 84–93. Retrieved 2025-08-12, from https://linkinghub.elsevier.com/retrieve/pii/S2211695813000160 doi: 10.1016/j.dcm.2013.04.001

Spitzberg, B., & Gawron, J. (2016). Toward Online Linguistic Surveillance of Threatening Messages. *Journal of Digital Forensics, Security and Law*. Retrieved 2025-08-12, from http://commons.erau.edu/jdfsl/vol11/iss3/7/ doi: 10.15394/jdfsl.2016.1418

Stemler, S. E. (2004). A Comparison of Consensus, Consistency, and Measurement Approaches to Estimating Interrater Reliability. *Practical Assessment, Research & Evaluation*, *9*(4). Retrieved 2025-08-12, from https://openpublishing.library.umass.edu/pare/article/id/1540/ (Publisher: University of Massachusetts Amherst) doi: 10.7275/96JP-XZ07

Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge University Press.

Swales, J. (2004). *Research genres: Explorations and applications*. Cambridge University Press.

Tompkinson, J. (2023). *Spoken threats from production to perception*. Cambridge University Press.

Upton, T. A., & Cohen, M. A. (2009). An approach to corpus-based discourse analysis: The move analysis as example. *Discourse Studies*, *11*(5), 585–605. Retrieved 2025-08-12, from https://journals.sagepub.com/doi/10.1177/1461445609341006 doi: 10.1177/1461445609341006

Xia, S. A. (2020). Genre Analysis in the Digital Era: Developments and Challenges. *ESP Today*, *8*(1), 141–159. Retrieved 2025-08-12, from http://doi.fil.bg.ac.rs/volume.php?pt=journals&issue=esptoday-2020-8-1&i=7 doi: 10.18485/esptoday.2020.8.1.7