languages

Special Issue Reprint

New Challenges in Forensic and Legal Linguistics

Edited by
Julien Longhi and Nadia Makouar

mdpi.com/journal/languages

MDPI

New Challenges in Forensic and Legal Linguistics

New Challenges in Forensic and Legal Linguistics

Guest Editors

Julien Longhi
Nadia Makouar

MDPI

Basel • Beijing • Wuhan • Barcelona • Belgrade • Novi Sad • Cluj • Manchester

Guest Editors
Julien Longhi
Institute of Digital
Humanities
CY Cergy Paris université
Cergy-Pontoise
France

Nadia Makouar
Laboratoire Praxiling
Université de Montpellier
Paul Valéry
Montpellier
France

Editorial Office
MDPI AG
Grosspeteranlage 5
4052 Basel, Switzerland

This is a reprint of the Special Issue, published open access by the journal *Languages* (ISSN 2226-471X), freely accessible at: https://www.mdpi.com/journal/languages/special_issues/WTIR4KZ312.

For citation purposes, cite each article independently as indicated on the article page online and as indicated below:

Lastname, Firstname, Firstname Lastname, and Firstname Lastname. Article Title. *Journal Name* **Year**, *Volume Number*, Page Range.

ISBN 978-3-7258-3931-5 (Hbk)
ISBN 978-3-7258-3932-2 (PDF)
https://doi.org/10.3390/books978-3-7258-3932-2

Contents

About the Editors

Julien Longhi

Julien is a Professor of Linguistics at CY Cergy Paris Université. He specializes in the discourse analysis of political and media texts, with a particular focus on ideologies, social media, and digital humanities. He has published books, articles, and edited volumes in the fields of pragmatics, semantics, and corpus linguistics, in addition to discourse analysis. He is the Director of AGORA lab and has several collaborations in the field of forensic linguistics.

Nadia Makouar

Nadia is an Associate Professor of Linguistics at the Université de Montpellier Paul Valéry. She specializes in corpus linguistics and the discourse analysis of political, media, and legal texts. Her research focuses on semantic analysis, modalities, and dialogism. Her publications include analyses and methodologies for approaching hate speech discourses, parliamentary debates, and lay legal language. She is part of the Praxiling Lab and has several collaborations in the field of legal linguistics.

Preface

Applied linguistics has experienced significant growth in the domains of justice, security, and law. The development of forensic and legal linguistics varies across legal contexts and depends on the relationships between universities and institutions, prompting critical considerations within applied linguistics. This Special Issue compiles pertinent contributions on the current state of forensic and legal linguistics, documents potential outcomes and contexts of study and collaboration, and emphasizes future challenges for the discipline. Researchers present recent advancements, particularly those linked to contemporary cases, explore new research avenues, and discuss the diversity of theories and methodologies employed. Various approaches—including corpus linguistics, NLP, discourse analysis, and pragmatics—are utilized, and examining their issues and implications from an applied perspective proves beneficial. This Special Issue encompasses thirteen contributions analyzing oral and written texts related to forensic and legal linguistics. The topics addressed include the analysis of legal discourse, the language of deception, comparisons of forensic texts, corpus-assisted analysis, the language of fraud and discrimination, and legal lay language. This Special Issue is addressed to researchers in the fields of applied linguistics, forensic linguistics, and legal studies. It is intended for scholars interested in the intersection of linguistics and law, justice, and cybersecurity.

Julien Longhi and Nadia Makouar
Guest Editors

languages

MDPI

Article

A Study of a Specialised American Police Discourse Genre: Probable Cause Affidavits

Audrey Cartron

Department of Applied Foreign Languages, Faculty of Foreign Languages and Cultures, Nantes University, CRINI, 44000 Nantes, France; audrey.cartron@univ-nantes.fr

Abstract: This paper focuses on the analysis of a specialised American police discourse genre and is based on a corpus of 115 probable cause affidavits. A probable cause affidavit is a sworn statement written by American police officers to state that there is probable cause to believe the defendant has committed (or is committing) a criminal offence and that legal action is required. After briefly presenting the methodological framework for this study, the paper intends to show how the police use specific linguistic, discursive and rhetorical strategies to serve a specialised purpose, which is to present the existence of probable cause to the relevant legal authorities. The findings indicate that officers use various discursive devices to inform but also—and perhaps more importantly—to convince their audience by means of a chronological and structured narrative of events that follows a prototypical three-fold internal organisation (exposition, investigation, resolution) signalled by specific linguistic markers. Finally, the paper intends to go beyond the objective description of events in order to highlight the assertive nature of this discourse genre and the additional rhetorical strategies used by PCA writers. It studies the emphasis placed on the expertise of the author, as well as the police classification of the offence and the progressive elaboration of the burden of proof.

Keywords: corpus linguistics; discourse analysis; English for Police Purposes; English for Specific Purposes; genre analysis; move analysis; probable cause affidavit; specialised discourse

Citation: Cartron, Audrey. 2023. A Study of a Specialised American Police Discourse Genre: Probable Cause Affidavits. *Languages* 8: 259. https://doi.org/10.3390/languages8040259

Academic Editors: Julien Longhi and Nadia Makouar

Received: 2 September 2023
Revised: 26 September 2023
Accepted: 9 October 2023
Published: 3 November 2023

1. Introduction

Due to the multiple interactions between police forces (specialists) and other members of society (non-specialists), English for Police Purposes (EPP) might intuitively appear less specialised (Petit 2010, §12) than Scientific English, for example. Nevertheless, EPP can be considered to be a specialised variety of English located at the crossroads of forensic and legal languages, with specific linguistic (Philbin 1996; Poteet and Poteet 2000), discursive (Johnson et al. 1993; Gaines 2011; Rock 2017) and cultural (Fielding 1994; Reiner 2000; Cartron 2023b) characteristics that deserve to be studied in depth. Among the various approaches that can be used to investigate specialised languages, genre analysis provides an interesting insight into the specialisation of the discursive community and its practices, taking into account both linguistic and extralinguistic features (Swales 1990, pp. 24–27; Beacco 2004, p. 116; Bhatia 2017, p. 6). As far as English for Police Purposes is concerned, this specialised variety of English is characterised by a diversity of genres, both spoken—such as police interviews, radio communications or court testimonies—and written—police reports, manuals or codes of ethics, for instance[1].

This paper focuses on the analysis of a specialised American police discourse genre belonging to the category of police reports and is based on a corpus of 115 probable cause affidavits (PCAs)[2] written by American police officers from different police forces (police departments, sheriff and county law enforcement agencies, as well as federal law enforcement agencies). In the United States, police officers are required by the Fourth Amendment of the Constitution to present probable cause and to justify that legal action is required:

> The right of the people to be secure in their persons, houses, papers, and effects, against unreasonable searches and seizures, shall not be violated, and no Warrants shall issue, but upon probable cause, supported by Oath or affirmation, and particularly describing the place to be searched, and the persons or things to be seized (Library of Congress n.d.).

In order to do so, officers write a probable cause affidavit[3], a sworn statement to state that there is probable cause to believe the defendant has committed (or is committing) a criminal offence and that the facts support the claim to make an arrest, conduct a search or seize the property (Crespo 2020, pp. 1279–80). Three different degrees of proof can be identified in the American legal system: reasonable suspicion, probable cause, and beyond reasonable doubt. Probable cause is the intermediate burden of proof and requires more evidence than reasonable suspicion (Taslitz 2010, p. 146) but less than beyond reasonable doubt. Therefore, it is an intermediate burden of proof between suspicion and certainty, and the police must gather sufficient evidence—both qualitatively and quantitatively—to support the hypothesis of the respondent's guilt.

Probable cause affidavits provide a brief summary of the events and identify the main parties involved, such as the victim(s), suspect(s) and witness(es). PCAs form a set of textual productions with a single communicative aim: to present the facts objectively to legal authorities (police superior, district attorney, judge or other actors in the judicial process). Based on the police officers' statements—as well as on other evidence and information from the case—the competent judicial authorities can validate (or not) the existence of probable cause; that is to say, they can determine whether there are grounds to believe the defendant has committed (or is committing) a criminal offence. Probable cause affidavits can be drawn up in several situations: either the individual has already been taken into custody, and the police must show the judge that probable cause exists to justify the legal value of the arrest, or this has not yet happened and the police must prove probable cause in order to ask a judge to issue an arrest warrant. The document can also be written as part of an application to a judge for a search warrant.

In the literature dealing with English for Police Purposes, several lines of enquiry relating to the discursive practices of police officers can be identified. Studies on police discourse tend to focus on specialised communication and practices as well as on major police discourse genres. They mainly deal with suspect interviews (Baldwin 1993; Leo 1996; Magid 2001; Haworth 2006; Benneworth 2009; Cartron 2023a), victim/witness interviews (Rock 2001; Milne and Bull 2006; Dando et al. 2009), police reports (Coulthard 2002), police calls (Tracy and Tracy 1998; Rock 2018), caution and Miranda warnings (Rock 2007; Heydon 2013), radio communications (Glaister 2006), interactions with professionals of related specialised fields (Johnson 2003; Charman 2013), police humour (Holdaway 1988; Gayadeen and Phillips 2016; Cartron 2023b), and *policespeak* (Fox 1993; Johnson et al. 1993; Hall 2008). However, according to the author's knowledge, no extensive and in-depth linguistic and discourse analysis has been conducted on the specialised American genre of probable cause affidavits.

After describing the methodological framework on which this study is based (Section 2), the paper provides a detailed move analysis of probable cause affidavits and shows how police officers introduce the existence of probable cause to the relevant legal authorities through the presentation of a chronological and structured narrative of the events (Section 3). The article then presents additional rhetorical strategies used by PCA authors and sheds light on the probative value of this discourse genre. It intends to go beyond the objective description of facts in order to highlight the emphasis placed on the expertise and reliability of the author, as well as the underlying progressive elaboration of the burden of proof (Section 4).

2. Materials and Methods

2.1. General Methodological Framework and Research Question

Several authors (Petit 2002; Wozniak 2011; Van der Yeught 2016; Stark 2020; Cartron 2022) have contributed to the development of a tripartite methodological protocol aimed at proposing descriptive characterisations of specialised varieties of English (SVEs). This three-fold protocol is based on the study of the discursive, linguistic and cultural features of the specialised domain under study. These three approaches are complementary and offer the possibility to present a holistic, structured and methodological description of specialised languages. The present study focuses on a discursive approach to English for Police Purposes and takes into account the linguistic and extralinguistic characteristics of the specialised language under study, including the context of production, the identity of the actors participating in the communicative event, the specific features of the specialised field as well as the aims of the communicative event (Charaudeau 2009, p. 41). Vijay Bhatia (1993, pp. 22–35) stresses the need to (re)place a discourse genre in situation and in context. A probable cause affidavit, for instance, cannot be analysed without taking into account the context in which it is produced, whether it is the immediate textual context (peritext), the context of reception (which may be reflected in the presence of the author and the addressee in the text, for example), the context of production (a particular offence, involving specific actors, at a given moment), or the social and cultural context, and more specifically the American judicial context (the concept of probable cause as specific to the United States, the legislation in effect in the state where the alleged offence was committed, etc.).

This paper deals with the detailed study of probable cause affidavits. Following the pioneering work of John Swales (1990), the rhetorical organisational patterns of PCAs were studied through a detailed move analysis of the genre. This approach consists of identifying the discursive or rhetorical units (called "moves") that perform a specific communicative function and serve the overall specialised purpose of the genre. In order to analyse a representative sample of the genre under study, it was decided to gather a corpus of authentic productions from American police officers. Two different approaches can be considered to investigate a corpus:

> "corpus-based" investigations, which are undertaken to check the researcher's intuition about language use, and "corpus-driven" investigations, where the researcher approaches the corpus data with an open mind to see what patterns emerge (Nesi 2013, p. 407).

The present study focuses on corpus-based investigations and concentrates on the following research question: how do police officers use specific discursive, linguistic—in terms of lexicon, phraseology and syntax—and rhetorical strategies in probable cause affidavits to serve a specialised purpose, which is to present the existence of probable cause to competent legal authorities? However, as it would be reductive to be limited by the rigid framework of a starting hypothesis (Martin 1997, §18), the author remained open to other leads or significant aspects that might emerge from the corpus during its exploration.

2.2. Overcoming the Lack of Accessibility of Sources

To study the specificities of EPP genres, it is necessary to gather authentic productions from police officers in order to undertake detailed and targeted analyses of specialised discourse. Indeed, the study of the discourses emitted by a given specialised community must be based on corpora composed of primary and authentic sources (Wozniak 2019, p. 5). However, in the field of EPP, collecting authentic materials—written and oral—produced by professionals is not an easy task (Oxburgh et al. 2010, p. 59). Internal productions within the professional police community can be confidential in order to guarantee the presumption of innocence[4], to protect victims and witnesses, to ensure the safety of police officers and their families and to avoid any effect on ongoing investigations. As Brodeur and Monjardet (2003, pp. 11–12) point out, in many countries, the legitimacy of police

secrecy is sanctioned by law. In the United Kingdom, for instance, data protection laws prevent many police documents from being made accessible to the general public.

In the United States, the Freedom of Information Act (1966) defends the principle of the right to information and makes it mandatory for federal agencies to hand over their documents to anyone who requests them. However, the legal obligation to make police documents accessible or not depends on the legislation in each state. In some states, journalists specialised in criminal cases, legal professionals, or even ordinary citizens can send requests for access to files on past or current cases. The Berkeley Graduate School of Journalism addresses the complex question of access to police documents in the United States and provides an online guide listing several sources that make authentic police records available to the public, including the American website *The Smoking Gun* (Grabowicz 2014). Created in 1997, this website belongs to the American group Turner Broadcasting System, a subsidiary of Warner Media, which runs, among others, the news channel *CNN*. The website is specialised in the publication of legal documents (Carr 2008), including police reports, arrest records and probable cause affidavits obtained through different sources: from government and law enforcement sources, via Freedom of Information requests, and from court files nationwide (The Smoking Gun 2020). Journalists of *The Smoking Gun* investigate criminal offences, and they publish police or court documents relating to these cases on the website[5]. PCAs studied in the present paper were selected from the *Smoking Gun* website.

2.3. Collecting and Analysing Data

Since the objective was not to study the diachronic evolution of probable cause affidavits over time, a synchronic perspective was adopted, and the collection of documents was limited to three years. Texts published on *The Smoking Gun* website between January 2018 and December 2020 were pre-selected. Among the 622 documents that were published, only those belonging to the category of probable cause affidavits and exclusively those for which it was possible to identify the date when it was written, as well as the author (or at least the corresponding police force), were included. Documents that were incomplete (missing pages) or unofficial (labelled "Unofficial document", "Unofficial copy", or "Not certified copy") were discarded. As the files available on the website were scanned versions of original documents, they were in image format (.jpeg files). They were then converted to text files (thanks to an optical character recognition software[6]) so that computerised analyses could be carried out using the corpus analysis toolkit AntConc (version 3.5.8.0). In order to make sure that the original and the converted texts were identical, each document was carefully proofread to correct the numerous missing, misspelt or truncated words and other typographical, linguistic or punctuation errors that were generated during conversion. Some texts were also entered manually when the conversion tool did not provide a usable result. These different steps led to the constitution of a corpus of 68,133 words, gathering 115 probable cause affidavits from 68 different American law enforcement agencies and from 18 different states. Although constrained by the question of the accessibility of the sources, the size of the selected corpus seemed adequate to study the process of specialisation at work in this specialised genre and to carry out quantitative and qualitative analyses jointly.

In order to study the multi-faceted genre of PCAs, a modular approach (Roulet n.d., p. 21) has been chosen based on the idea that a genre can be considered as a system combining various smaller parts called "modules". The lexical module looks at lexical units (nouns, adverbs, verbs, adjectives, pronouns) and the use of specific vocabulary or terms. The phraseological and syntactic module studies collocations, fixed phraseological units and, more generally, the relations between linguistic units (the use of active and passive voices or indirect discourse, for example). The structural module covers the formal characteristics of the genre, its internal structure (rhetorical moves) and its external structure (paratext). The combination of these three main modules leads to the accumulation of

knowledge on the specificities of the genre and, beyond that, of the specialised variety being studied.

Several tools and methods were used to study the various elements within each module. Firstly, careful reading and manual analyses of the selected texts were carried out throughout the process of collecting the corpus. Specific attention was paid to the lexical, phraseological, syntactic and structural characteristics of the genre. Detailed move analyses were also performed by the author on five probable cause affidavits from different US states and types of police forces. The procedure used to study discourse moves in PCAs included understanding the overall rhetorical purpose of the texts, identifying the different text segments as well as their function and purpose, and then studying and coding common functional and/or semantic themes represented by the various text segments (Kanoksilapatham 2007, p. 33). Secondly, this first-hand qualitative approach was supplemented by more quantitative and computerised processing of the data (Banks 2016, §33) to complete the characterisation of this discourse genre. As underlined by Budsaba Kanoksilapatham, a corpus-based analysis allows "for more complex and generalizable research findings, revealing linguistic patterns and frequency information that would otherwise be too labor intensive to uncover by hand" (Kanoksilapatham 2007, p. 36). For this study, the AntConc concordance was chosen because it offers the possibility of easily studying the behaviour of a word in context (*keywords in context* feature), as well as its distribution and place in each text of the corpus (*concordance plot*). It also allows us to identify the most frequently used words in the corpus (*word list*) and to single out collocations or compound terms (*clusters/n-grams*). Finally, some authors also advocate an ethnographic approach to the genre, which involves the validation—or invalidation—of analyses by a specialist in the field (Bhatia 1993, pp. 22–35). The genre of probable cause affidavits—including their content, aim, and purpose—was thus discussed with American (as well as some British) police officers who were interviewed between December 2019 and March 2022. The following sections of the article describe the findings of this study on the characterisation of probable cause affidavits.

3. Move Analysis of a Chronological and Structured Narrative of Events

3.1. A Three-Fold and Prototypical Internal Structure

The internal structure of probable cause affidavits is not fixed, but several regularities emerge. Following John Swales's genre analysis approach (Swales 1990), the results of the present research indicate that PCAs generally adopt a prototypical structure characterised by three rhetorical moves signalled by specific linguistic markers. It is organised around a chronological narrative of the facts, and the three moves follow a prototypical narrative structure in three acts: (1) exposition, (2) investigation, and (3) resolution/denouement. The first move is devoted to the presentation of the initial context. After specifying the exact date, time and location of the intervention, the police officer presents the triggering event or inciting incident (emergency call, *flagrante delicto* observed during a patrol, transfer of a file between two police units or forces . . .) and the type of offence under investigation. In the second move, the police officer sheds light on the various investigative steps taken (such as taking statements or viewing recordings from the cameras that filmed the scene) and the evidence collected. In the third and final move, the conclusion is made up of the findings on the existence of probable cause and details of the arrest of the suspect when applicable. Therefore, the reader is guided step by step by the author, who presents the events in a chronological and structured manner. The breakdown of affidavit PC_LA_WestMonroePD_2019[7] (Figure 1) illustrates the use of this prototypical three-stage structure[8].

The actions of the police are not always explicitly mentioned. Some affidavits present a chronological narrative of the offence itself, with an omniscient point of view and a focus on the actions of the suspect rather than on the investigation. However, the elements presented in the document are similar: initial context, inciting incident, implicit presentation

of the investigation, and collection of evidence. In PCAs, each move has specific linguistic characteristics, as exemplified in the following subsections.

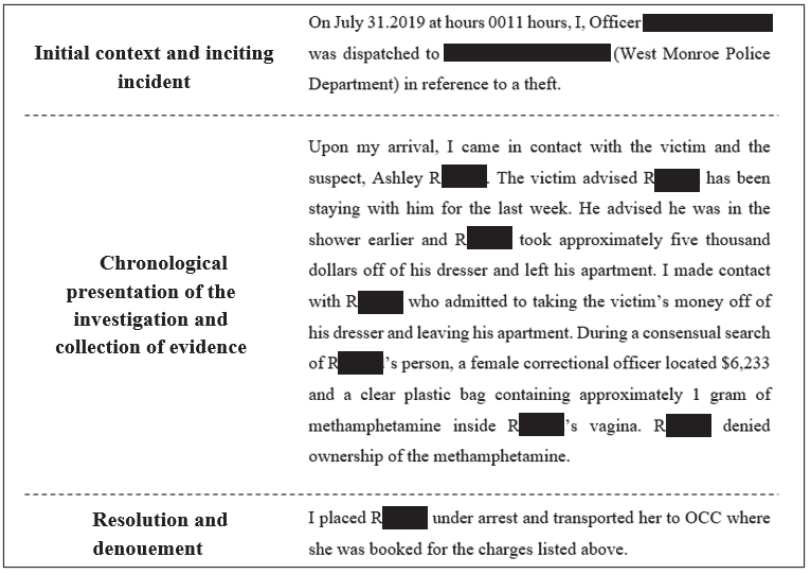


Figure 1. Prototypical structure of probable cause affidavits: example and breakdown of affidavit PC_LA_WestMonroePD_2019.

3.2. Examples of Linguistic Markers for Move 1 (Exposition)

The first words of probable cause affidavits always set the facts in a precise temporal and geographical context. The dates, times and places of police intervention are the first elements mentioned, generally followed by the type of offence, plunging the reader *in medias res* in the recounted events. This pattern is recurrent in police reports, as an American police officer underlined:

> There definitely is [a police-style of writing]. When it comes to police officers or Detectives writing reports, sure, it's a definite style. It's very mechanical. There isn't a lot of fluff. It usually starts out on the day, date and time. So, "On Thursday, May 4th, at about eleven ten a.m., myself, Sergeant [*states his own name and surname*], on Squad 21 15 observed . . .", then you go into whatever the story is (date of the interview: 4 June 2020).

Move 1 of PCAs is characterised by the extensive use of contextual linguistic units (adverbs, prepositions, prepositional phrases, verbs, etc.). For instance, the locations and types of incidents are presented with specific recurring linguistic markers (Table 1).

Table 1. Recurrent linguistic markers in the first rhetorical move of probable cause affidavits.

Linguistic Markers Introducing the Location of the Intervention/Incident	Linguistic Markers Introducing the Type of Incident
responded to (40 occurrences) *was/were dispatched to* (18)	*for/on a report of* (11 occurrences) *in reference to* (32) *responded to* (3) *was/were assigned to* (2) *was/were dispatched to* (2)

The contexts of the use of these markers and their distribution in PCAs indicate that they form a linguistic specificity of the first rhetorical move. For instance, Figure 2 is

a screenshot of the Concordance plot tool of AntConc, showing the distribution of the prepositional phrase *in reference to* (32 occurrences) in different texts[9]. It demonstrates that the item is used extensively and exclusively at the beginning of affidavits to introduce the type of offence.

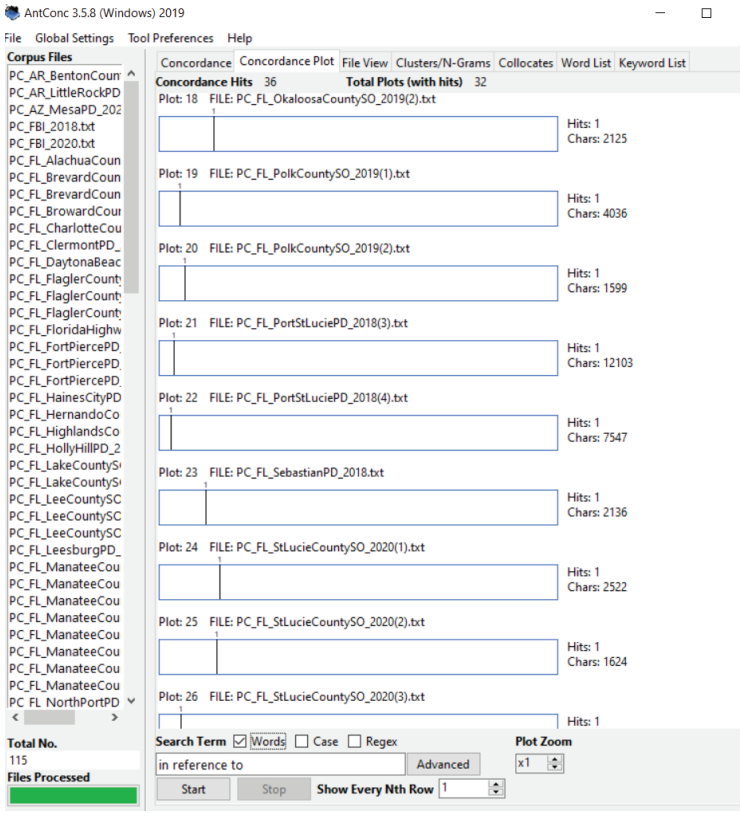


Figure 2. Distribution of the prepositional phrase *in reference to* in probable cause affidavits (AntConc, Concordance plot).

Furthermore, Figure 3 illustrates the contexts of use of this prepositional phrase and shows that it collocates with nouns designating generic categories of incidents. These nouns may be legal terms given to the offence in the law (*theft*, *battery*) or, in most cases, a much vaguer classification (*disturbance*, *sexual offence*, and even *suspicious incident*).

Interestingly, this initial description of the offence reflects the information given to the police officer when they are assigned to the case and reveals their initial imprecise knowledge of the facts when they are dispatched. The investigation then enables the classification of the criminal offence more precisely, as it will be discussed in Section 4.2.

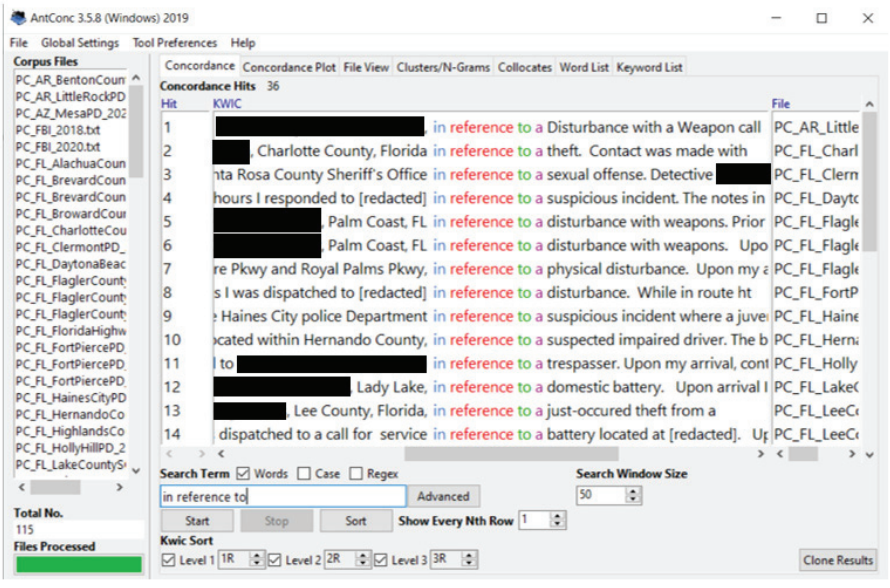


Figure 3. Concordance lines for *in reference to* in PCAs (AntConc, KeyWord In Context).

3.3. Examples of Linguistic Markers for Move 2 (Investigation)

In the second move, the different steps of the investigation are precisely traced using various temporal markers such as *after*, *before*, *during*, *hours*, *later*, *then*, *time*, *when* or *while*. It shows the need for a very thorough description of the facts. The frequent use of *approximately* (161 occurrences) was deemed intriguing as it seemed to contradict the emphasis on precision specific to the writing of probable cause affidavits. However, a study of the contexts in which this adverb of approximation was used revealed that it often collocates with extremely specific temporal details, such as *at approximately 0124 h*, paradoxically reinforcing exhaustiveness.

Moreover, law enforcement representatives talk to a wide range of people: victims, suspects, witnesses, and other specialists (forensic experts, police colleagues present at the crime scene or previously in charge of the case). Each protagonist is clearly identified using categorising nouns in order to establish the agents of the various actions reported. The significant use of *defendant* and *victim* can be highlighted, as they are the first two most frequently used common nouns in the corpus, with 460 occurrences (rank 18) and 369 occurrences (rank 22), respectively. The regular use of the nouns *officer(s)* (196 times), *deputy* (118), *police* (99), and *affiant* (69) can also be highlighted. Finally, third-person pronouns are also numerous, and *he*, *him*, *his*, *she* and *her* are among the twenty most frequent words in the corpus. Additionally, investigative acts carried out by the police are also clearly identified and indicated by the use of verbs such as *observed* (139 occurrences), *asked* (122 occurrences), *made contact with* (42 occurrences) or *spoke to/with* (43 occurrences). To this extent, probable cause affidavits provide insight into the practices of the specialised community. For example, the following extract from an affidavit drafted by a Florida police officer illustrates the procedure to be followed in the event of suspected drunk driving:

> Deputy S arrived on scene and assisted with demonstrating the Standardized Field Sobriety Exercises. Deputy S explained the horizontal gaze nystagmus exercise to the defendant and he replied he understood the instructions given. [. . .] Deputy S asked him multiple times to only follow the tip of the pen with his eyes and reminded him not to move his head. The defendant continued to move his head [. . .]. Deputy S then explained and demonstrated the walk and turn exercise to the defendant. The defendant was unable to stand in the

> heel to toe position without losing his balance [. . .]. Deputy S then explained and demonstrated the one leg stand to the defendant. [. . .] The defendant then stood with his feet next to each other without lifting a foot up. The defendant was reminded to pick a foot of his choosing to complete the exercise. [. . .] The defendant raised his foot for approximately half of a second before losing his balance and setting his foot down. [. . .] After my investigation I determined the defendant was under the influence of an alcoholic beverage and operating his golf cart under the influence of alcohol (PC_FL_SumterCountySO_2020(1)).

Therefore, PCAs depict—whether explicitly or implicitly—the gestures, practices and procedures and the day-to-day life of an American policeman in the field.

Last but not least, the extensive use of indirect discourse and reported speech verbs can also be highlighted. Among the hundred most frequently used words of the corpus, the following verbs were identified: *stated* (rank 20, 420 occurrences), *advised* (rank 42, 178 occurrences), *said* (rank 51, 155 occurrences), and *told* (rank 63, 123 occurrences). The preterit *stated* is used frequently; it is the twentieth most used word and the second most used verb (after *was*) in the corpus. It appears in 82 of the 115 texts and is frequently used several times in the same document, as shown in Figure 4. This recurrence can be explained by the fact that the derivative *statement* refers to words declared before a police officer and intended to be produced in court.

The occurrences *stated* are spread throughout probable cause affidavits and signal the use of reported speech and the presentation of information obtained during the various interviews conducted during the investigative work. The syntactic rule provides for the adaptation of pronouns when using indirect discourse, but some errors were identified in the corpus, remains of an incomplete transition from direct to indirect speech, as in the following example:

> While sitting in the turning lane on Highway 27, the defendant told the victim to get out. The defendant stated the police will find *you* a new home (our italics, PC_FL_HainesCityPD_2019).

Interestingly, despite the wide variety of words belonging to the class of declarative verbs, *state*, *advise*, *say* and *tell* are selected as priorities by affidavit writers. This lexical preference for a restricted spectrum of verbs is corroborated by the few occurrences of verbs with similar semantic characteristics. For example, the verbs *added* (1 occurrence as a verb of declaration introducing reported speech), *explained* (20 occurrences), *indicated* (7 occurrences), *mentioned* (1 occurrence) or *reported* (10 occurrences) are very rarely used. Other variants are never used in the corpus, such as the verbs *declared*, *highlighted*, *underlined* or *pointed out*. This lack of variation in the formulations seems to indicate that the facts presented in PCAs take precedence over the form, as the writing is mostly motivated by a concern for concision, brevity, clarity and efficiency. Moreover, this low vocabulary richness also suggests that PCAs are very formulaic in nature. This genre is frequently written by police officers, and, as a result, lexical choices, as well as collocations, become fixed and routinised.

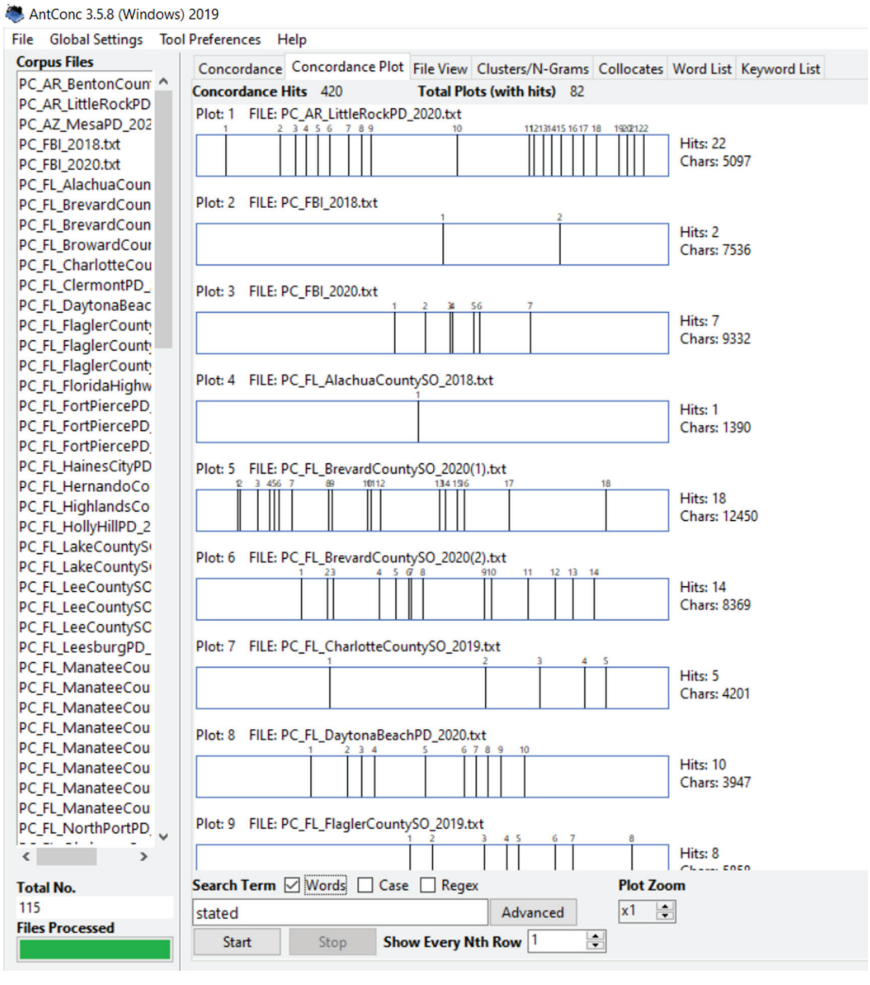


Figure 4. Distribution of *stated* in PCAs (AntConc, Concordance plot).

3.4. Examples of Linguistic Markers for Move 3 (Resolution)

Move 3 concludes probable cause affidavits and presents, in a few words, the police officer's conclusions following the investigation they have conducted. The concluding elements differ from one police force to another, and there are many variations in this rhetorical move. In some documents, the author stresses that the evidence gathered establishes the existence of probable cause. The officer indicates that all the elements are present to observe a breach of the law and precisely designates the offence(s) committed and the corresponding legal text(s). Certain lexical elements are specific to this rhetorical move. The pattern <*Based on* [evidence], *probable cause* . . .> is used several times (25 occurrences), as in the following example:

> *Based on* the above facts, statements and physical evidence provided, *your Affiant has probable cause to believe* and does believe that the above listed probable cause, all lead to the substantiation that defendant, S, has committed a violation of the laws of the State of Florida, to wit: Solicitation to commit 1st degree Murder, contrary to section 777.04 (4-B), Florida Statutes and Solicitation to commit an Occupied Burglary with a Battery, contrary to section 777.04 (4-C) (our italics, PC_FL_BrevardCountySO_2020(1)).

The evidence referred to in the conclusion ("the above facts, statements and physical evidence") are relatively vague categories with anaphoric value, as they refer to the evidence previously referred to. In addition, at the end of the affidavit, some authors highlight the actions taken by the police to close the case, and more specifically, the arrest of the respondent, as shown by the last words of this probable cause affidavit:

> Based on my observations on scene, *I took M into custody* for FSS 784.045(1A1)—Aggravated battery for striking the victim on the head with the can of Spaghetti's. *M was transported to St Lucie County Jail without incident. This case was Cleared by Arrest* (our italics, PC_FL_StLucieCountySO_2020(2)).

Therefore, probable cause affidavits are chronological narratives of the facts, structured in three stages, and each move of this prototypical structure serves an overarching communicative purpose (Bhatia 1993, p. 37): to inform and guide the reader but also to convince legal authorities of the existence of probable cause. In the course of this study of probable cause affidavits, it became clear that the communicative aim of PCAs is not only to present a series of facts objectively but also to model the discourse in order to serve a specialised purpose and, more broadly, to provide usable content for the judicial process. The last section of this article argues that police officers use specific discursive procedures to inform, but also—and perhaps above all—to convince and persuade the reader(s) of the guilt of the individual, and not just of its probability.

4. Additional Rhetorical Strategies: From Probability to Certainty?

4.1. The Author's Expertise and Credibility

In police reports, discourse modelling is motivated by the underlying desire to convince the reader of the veracity of the facts presented. As in academic genres studied by Ken Hyland (2005, pp. 173–74), police reports are written with the aim of persuading the reader by using various rhetorical techniques, including the credible representation of themselves, their actions and the events observed:

> [A]cademics [are] not simply producing texts that plausibly represent an external reality, but also as using language to acknowledge, construct and negotiate social relations. Writers seek to offer a credible representation of themselves and their work [and] controlling the level of personality in a text becomes central to building a convincing argument. Put succinctly, every successful academic text displays the writer's awareness of both its readers and its consequences (Hyland 2005, pp. 173–74).

In PCAs, the expertise of the author is sometimes explicitly presented. In some police forces, affidavits begin with an introductory paragraph that briefly describes the officer's career: number of years of service, skills acquired during various training courses, types of cases handled, etc. In the State of California, this introductory paragraph is informally referred to as the "hero sheet":

> The way that we write our affidavits in the State of California usually starts with what we jokingly refer to as the hero sheet. We explain to the judge who we are, and when we're forming our affidavit, we refer to ourselves, the person that is swearing to the facts and circumstances that we're in this affidavit, as we are seeking this search warrant. We refer to ourselves as the affiant, or sometimes people will pronounce it as affiant. So, in that hero sheet section of the affidavit at the beginning I explain my training and experience, because later on in the affidavit I'm going to ask the judge to take my expert opinion into account when I sum up the meaning of all those facts and circumstances, and what they mean as I lay out the basis for probable cause (Richardson 2018).

Several examples of this explicit presentation of the author's expert status were found in the PCA corpus, as shown by the following extracts from a police officer in North Dakota and from an FBI agent:

> I Detective J, attest to the following: That *I am a trained and licensed Peace Officer* with *9 years of experience* with jurisdiction to enforce state law in city of Bismarck, Burleigh County, North Dakota. In 2009, *I successfully completed Military Police Academy for the United States Marine Corps* in Fort Leonard Wood, MO. In 2010, *I attended the Devils Lake Regional Police Academy* and was hired by the Mandan Police Department in 2010. In 2013, I was hired by Bismarck Police Department and currently work as an Investigator in the Investigation Section. *I have attended The Basic Course of Criminal Investigation by BCI, The Reid Investigator Interview and Advance Interrogation and Evidence Based Interrogation by the CTK Group. I have attended the National Fire Academy and taken Fire Investigation Essentials to Origin and Cause.* I have *over 1300 h of Law Enforcement related training* (our italics, PC_ND_BismarckPD_2019).

> I am *a Special Agent with the Federal Bureau of Investigation* (FBI) within the United States Department of Justice and have been so employed since March 2000. I *primarily work in the Minneapolis, Minnesota division.* Prior to my employment with the FBI, *I served as an Indiana State Trooper for approximately 3 years.* As a Trooper *my duties included criminal investigation, traffic offenses, and gaming regulation.* During my tenure with the FBI, *I have actively participated in investigations, including violent crimes* in Indian County *and international terrorism.* Since 2009, I have been *the Minneapolis Division Weapons of Mass Destruction Coordinator* and have *experience investigating explosives.* I have *a Bachelor's Degree* from Indiana University (our italics, PC_FBI_2020).

Various elements are mentioned, including the author's status, training institution(s), number of years of experience, previous places of practice, current assignment and various training courses received. This accumulation actively contributes to the construction of the author as a credible expert in the field of law enforcement. To a certain extent, the facts relating to the offence subsequently stated are difficult to contest because they are backed up by credentials and in-depth professional expertise. In PCAs, police officers elaborate and structure narratives by utilising the "rhetorics of reality" and, more specifically, the "reality production kit" evoked by Alexa Hepburn (2003, p. 181). For instance, the authors foster category entitlement by "construct[ing] [their] talk as coming from a category that is credible or knowledgeable in a way that is relevant to the claim" (ibid.). Additionally, expertise is recognised by the courts as they rely on the training and experience of police officers to assess probable cause:

> [T]he [Supreme] Court has been reasonably consistent in explicitly stating, or at least assuming, that a police officer's training and experience help support the existence of probable cause and reasonable suspicion. And the lower courts have followed suit (Kinports 2010, pp. 752–54).

In order to establish the existence of probable cause, police officers must rely on their expertise and knowledge regarding legal definitions of offences, well-known local criminal characters, different types of *modus operandi*, various investigative approaches and interview techniques (South Carolina Law Enforcement ETV Training Program 1976b, pp. 19–20). Thanks to their specialised knowledge, police officers can, for example, interpret certain facts or statements made by defendants:

> Shortly thereafter, an explosion is audible in the video and R repeatedly yelled "good shot my boy" and "Fuck 12." *I know from my training and experience that* the term "Fuck 12" is a derogatory phrase often directed at law enforcement officers (our italics, PC_FBI_2020).

> I spoke with Z. Z said he does use "dabs". *I know from my training and experience that* dabs is a commonly used name for hashish oil (our italics, PC_ND_MandanPD_2018).

In these two extracts, the authors use their police knowledge to explain the two terms to the lay reader(s) in order to secure the understanding of the meaning of these statements

used by defendants (a pejorative expression to refer to the police in the first example and the designation of illegal drugs in the second one).

Moreover, the authors' credibility and seriousness and their status as a bearer of truth are also enhanced by the fact that they officially take an oath and declare on their honour the truthfulness of the narrated events. PCAs are sworn statements made in writing before a competent authority (notary public, deputy clerk of the court, assistant state attorney, magistrate or certified officer). This is indicated by the words *Before me* in the sentences "Before Me, the undersigned authority, personally appeared [name of police officer] . . ." or "Subscribed and sworn to (or affirmed) before me" at the beginning or end of documents. When signing an affidavit or sworn statement, the police officer solemnly declares that the stated facts are true. This is reflected in the use of frequently used fixed phraseology such as "The undersigned certifies and swears that . . ." or "I swear that the above statement is correct and true to the best of my knowledge and belief". In the event of perjury—lying or giving false evidence—police officers are liable to severe penalties, including dismissal, redundancy or imprisonment. John Michael Callahan, deputy sheriff for Plymouth County (Massachusetts) and former NCIS and FBI special agent, outlines the consequences of deliberate misrepresentation and omission in affidavits drafted by American officers:

> Carlos Luna, a Boston Police Department (BPD) Detective, obtained a search warrant for a residence based upon his sworn affidavit. Luna's affidavit claimed he received information from an informant that illegal drug activity was occurring at that residence. Luna and other officers went to the residence to execute the warrant. During a forced entry, shots were fired from inside the residence and an officer was killed. Albert Lewin was charged with murder of the officer. During legal proceedings that followed, Lewin's lawyer moved for disclosure of Luna's confidential informant. The judge granted the motion, but the prosecution was unable to produce the informant. As a result, the trial judge dismissed the Lewin indictment. Detective Luna submitted a new affidavit in an effort to obtain reinstatement of the charges against Lewin. Luna admitted to making substantial material misstatements in his search warrant affidavit including the facts that he attributed to his informant. The case against Lewin was reinstated by the Massachusetts Supreme Judicial Court, but Lewin was later found not guilty of the officer's murder at trial. Detective Luna was subsequently charged and convicted of perjury and filing false police reports (Callahan 2019).

The authors' specialised knowledge, their position within the specialised community and the action of oath-taking are elements that guarantee and reinforce the seriousness and reliability of the facts narrated in probable cause affidavits. Additionally, the expertise and credibility of the authors also legitimise the signposting work they perform when classifying the offence, thus laying the foundations of the judicial process.

4.2. Signposting and Classification of the Offence

When they look at the documents in a case file, actors involved in the judicial system must be able to quickly identify the type of case presented and, in particular, the category of the offence. Therefore, the police carry out an operation of signposting, which consists of classifying the case in one (or more) specific category(ies) of criminal offence(s). This initial classification conditions the reception of the text as a whole, as it orients the case towards a defined legal framework and, consequently, towards the nature of the expected evidence. To follow the metaphor of the railroad switch on a railway, the author of a probable cause affidavit drives the case in the direction of one or more common law precedents and directs it towards legal lines of final destination that have been determined over the decades by case law: "legislators codify offences *ex ante*, and [. . .] police and prosecutors confine their collective attention to the catalogue of what has already been defined as criminal" (Bowers 2014, p. 997). By classifying the offence, American police officers attempt to insert the facts into the wider context of the legal system. In order to do so, the police specifically name the offences that were committed and refer to the corresponding legislation. This

aspect is illustrated by the use of specific legal terminology and, more precisely, fixed phraseological units both in the peritext (Figure 5, example 1) and in the body of the text (Figure 5, example 2).

PC_FL_DaytonaBeachPD_2020

CHARGES	DOMESTIC VIOLENCE? Yes ☒	Attachments: Affidavit(s)? ☐	Statement(s) ☒	NTA Schedule ☐	Report ☒	Traffic Infraction(s) ☐	DUI ☐	Total Charges: 2
#1	Charge: Child Abuse w/o Great Harm	FEL ☒ MISD ☐ ORD ☐	FS/ORD: 827.03(1)	Citation No.:	Bond: No bond			
#2	Charge: Agg Asslt w/Deadly Weapon w/o Intent-Kill	FEL ☒ MISD ☐ ORD ☐	FS/ORD: 784.021(1)(A)	Citation No.:	Bond: No bond			
#3	Charge:	FEL ☐ MISD ☐ ORD ☐	FS/ORD:	Citation No.:	Bond:			

PC_FL_BrevardCountySO_2020(1)

AFFIDAVIT FOR ARREST WARRANT

State of Florida
County of Brevard

BEFORE ME [redacted] a sworn law enforcement officer, personally came Agent [redacted], of the Brevard County Sheriff's Office, who being duly sworn deposes and says: that Affiant has reason to believe and does believe that probable cause exists for the arrest of [redacted], **Date of Birth** [redacted], **Social Security Number** [redacted], **last known address of** [redacted], **5'2" 160lbs, Black Female** for a violation of the laws of the State of Florida, to wit **Solicitation to commit 1st degree Murder,** contrary to section **777.04 (4-B),** Florida Statutes **and Solicitation to commit an Occupied Burglary with a Battery, contrary to section 777.04 (4-C)** which occurred at [redacted], **Brevard County, Florida, 32904.**

Figure 5. Classification of the case by the police in PCAs.

Therefore, the type of criminal offence(s) is clearly stated. It can easily be identified by a reader who is unfamiliar with the case, as the terms used by the authors reflect how offences are referred to in legal texts: *child abuse without great harm* and *aggravated assault with a deadly weapon without the intention to kill* (example 1 above), a *solicitation to commit first-degree murder* and *solicitation to commit an occupied burglary with a battery* (example 2).

As pointed out in Section 3.2, the first designation of the offence in probable cause affidavits is not always based on a precise classification because the account reflects the imprecise initial knowledge of the facts available to the police officer when assigned to the case. The investigation then enables the classification of the criminal offence more precisely, and this progression is sometimes perceptible. For example, the relatively vague reference to *a sexual offence* at the beginning of affidavit PC_FL_ClermontPD_2020 is then classified more narrowly when the police officer uses precise legal terms: *Lewd or Lascivious Exhibition in Violation of Florida State Statute 800.04 7(a)1*. Similarly, in PC_OK_RogersCountySO_2018, the first reference to the offence is *having sex with a pony,* and it then becomes *Indecent Exposure and Bestiality* because the incident is associated with a specific and defined legal framework. Therefore, good knowledge of common law precedents and legal texts is an essential prerequisite for the authors. Police officers need to be familiar with existing legal frameworks, but they also need to continually update their knowledge because the definitions given to offences in legislative texts may evolve insofar as adaptations and modifications are necessary when a particular context arises. For example, the COVID-19 pandemic led to the implementation of new legislation (lockdowns, various bans and restrictions, border closures, etc.). In this context, an individual who deliberately coughed on a shop assistant (to protest against social distancing measures) was detained for aggravated assault:

> Based on the verbal/Written statements obtained on scene, Deputy C charged C with aggravated assault, given C intentionally and unlawfully threatened, by word or act, (coughing on) to do violence to P. At the time the threat was made (during the COVID-19 pandemic), C appeared to have the ability to carry out the threat, by active coughing on P. C's threat created in the mind of P a well-founded

> fear that the violence was about to take place, and assault was made either with a deadly weapon or with a fully formed conscious intent to commit a felony (PC_FL_VolusiaCountySO_2020).

Finally, the police officers' classification of the offence is not always definitive, as it may be re-classified later in the judicial process in light of the evidence provided by investigations. Therefore, the communicative aim at work in PCAs is to construct a modelled discourse that can be correctly interpreted within a given context of jurisdictional precedents. Police officers' operation of signposting is reflected not only in the initial classification of the offence but also in the progressive construction of the burden of proof.

4.3. The Progressive Elaboration of the Burden of Proof

In PCAs, the burden of proof is built up through the accumulation of evidence. Police officers select from the wide range of information they receive and give priority to the decisive, even incriminating elements: "Probable cause is built like a stack of blocks—by piling one fact indicating guilt on top of another" (South Carolina Law Enforcement ETV Training Program 1976a, p. 9). The combination of verbal and physical evidence reinforces the probative force of the elements presented by the author. Additionally, writers of probable cause affidavits also diversify and multiply the sources of information: statements from the victim(s) and witness(es), interviews with suspects, evidence gathered by peers (police officers, scientific experts . . .), observations made at the scene of the incident, video-surveillance, etc. The progressive elaboration of the burden of proof is illustrated by the following PCA, in which a police officer interviews the victim and collects verbal and physical evidence:

> *I asked N to explain to me what happened*. N stated that he was bagging B's groceries and B got upset because he didn't like the way he was putting his chips into the bags. N stated after the groceries were bagged and the bill was paid B started to walk away. B then turned around and approached him and stated "Do you have a problem with me, because I have a problem with you". N then thinking that B was joking with him stated "do you?". [. . .] Then B quickly moved in N's direction and grabbed N by the throat/neck area and pushed him back against the register. [. . .] N then showed me where B placed his hand around his neck/throat. *I did observe there to be a dark red area to N's neck/throat. The area did look as it was turning to bruising. I did photograph this as evidence.* [. . .] *I asked N to provide me a written statement of the incident*, which he agreed to. *This incident was caught on the store video system. I reviewed the footage and did find that B in fact did grab/strike N in the throat area and pushed him up against the register* (our italics, PC_PA_FairviewTownshipPD_2019).

This extract exemplifies the use of two discursive and rhetorical strategies from the "reality production kit" (Hepburn 2003, p. 181). Corroboration and consensus (narrative corroborated by a witness/the victim), as well as active voicing (quotations to present supporting views), are used by the author to construct their arguments. In some cases (as in the above example from PC_PA_FairviewTownshipPD_2019), the reader can easily reconstruct the dialogue with one or several interlocutor(s). However, on many occasions in the corpus, the role of the enunciator disappears in order to place the emphasis on the collected statements and their content. The different steps of the investigation then become implicit:

> J stated P came into the office with regards to questions about the property. P started talking about a football game which led to a conversation about Collin Kaepernick. Conversation became heated and P became confrontational and threatening towards J (PC_FL_PortStLuciePD_2018(1)).

In such cases, discourse is modelled so that the questions asked by the investigators disappear in order to give primacy to the statements and evidence. Some affidavits are even characterised by a disappearance of the officers' actions in order to encourage the reader to

concentrate on the description of the facts relating to a breach of the law. This rhetorical strategy leads to the production of affidavits centred almost exclusively on account of the suspect's actions during the commission of the offence. This focus is adopted by various documents, including PC_UT_LaytonPD_2019:

> On 11/7/19 a male later identified as V ordered food from McDonald's inside of Layton Wal-Mart at anonymous-address. V then left with his food. V was wearing a dark blue sweater and blue jeans. V later returned to McDonald's and went behind the front registers into the employee area where customers are not allowed. V then proceeded to assault an employee at the register with his fists hitting the employee in the face. V then walked further back in the business into the kitchen area and assaulted another employee with his fists hitting the employee in the face as well. V then is heard saying you got my order wrong. The event was captured on surveillance cameras. V was identified by another officer on the Davis Crime Bulletin (PC_UT_LaytonPD_2019).

This affidavit mainly presents the facts that occurred and the temporality of the investigation disappears in favour of the temporality of the offence. As a result, readers of the affidavit, that is to say, outsiders who were not present at the scene, cannot measure the way in which the police officer guided, or even influenced, the exchange and the type of evidence gathered (Komter 2001, p. 368).

To put it in a nutshell, the aim of PCAs is to convince the competent judicial authorities to validate the existence of probable cause. In order to do so, police officers provide a modelled narrative of the facts and of police actions. As in most of the reports they write, police officers are not required to explicitly present a subjective analysis of the facts, and the aim is to convince by recounting events and presenting them following specific discourse conventions. Therefore, when writing police reports, police officers are part of a hybrid temporality because they are looking both to the past—events that have taken place—and to the future, as the documents will then be used in the judicial process and the future reception of the text by the reader(s) needs to be taken into account.

5. Conclusions

To conclude, when drafting probable cause affidavits, the police must gather sufficient evidence—both qualitatively and quantitatively—to justify the existence of probable cause and, ultimately, to support the hypothesis of the respondent's guilt.

Each move of the three-act prototypical structure of PCA described in Section 3.1 is meant to serve this mechanism. Indeed, the chronological and structured narrative of events is designed to persuade readers by presenting the facts in a logical, rational and coherent way. Move 1 (presentation of the context) places the case in a specific time and place, and the triggering event exposes a problematic situation that justifies police intervention. The second move (presentation of the investigation) enables legal authorities to assess the quality and quantity of the gathered evidence. Finally, in the last rhetorical move, the author evokes the details of police actions conditioned by the existence of probable cause (the arrest or the application for a warrant) and the resolution of the case. By rationally presenting a logical sequence of events (as in a demonstration), the author uses *logos*, one of the three rhetorical modes of persuasion defined by Aristotle—along with *ethos* and *pathos*—in his work *Rhetoric* (Chiron 2007). *Logos*, or persuasion through discourse, consists in showing that something is true or appears to be true, and this is precisely the aim of police officers when they present the details of the case in a coherent, chronological and structured narrative.

Furthermore, it can be argued that the shift from probability to certainty is also reinforced by the emphasis placed on the expertise and credibility of the author. This rhetorical strategy, referred to by Aristotle as *ethos*, is related to persuasion by character and consists in making the speaker worthy of belief through discourse. PCA authors present themselves as credible experts in the field of law enforcement, taking an oath before a competent authority and putting forward the specialised knowledge they acquired through

training and experience. Last but not least, the facts stated in probable cause affidavits participate in the initial classification of the offence, which can have a long-lasting impact on the case, and the burden of proof is progressively built. Once again, contrary to what the term *probable* suggests, there is no room for probability, doubt or uncertainty in probable cause affidavits, as the narrative does not highlight the probable dimension of the narrated facts but rather posits their veracity.

The present paper intends to contribute to characterising a barely-studied specialised language—English for Police Purposes—by providing an extensive and in-depth analysis of probable cause affidavits. It sheds light on the underlying communicative goal, as well as on the rhetorical moves and strategies that define this discourse genre, thus allowing a better understanding of the linguistic conventions and practices of American police professionals. It is hoped that these findings will be of interest to practitioners but also teachers and learners of police English, as well as to researchers characterising specialised varieties of English. Several lines of enquiry regarding police discourse remain open for future research, such as detailed and comparative studies of corresponding or related documents written by law enforcement officers from other English-speaking countries or in-depth analyses of other EPP genres (both spoken and written). Police language constitutes a promising and multi-faceted object of study that remains, for the time being, a relatively uncharted research territory in the ESP community.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board and Ethics Committee of Nantes University (protocol code 06102023) on 6 October 2023.

Data Availability Statement: The data presented in this study (i.e., extracts from probable cause affidavits written by American police officers) are available at http://www.thesmokinggun.com/documents, (accessed on 9 January 2021).

Conflicts of Interest: The author declares no conflict of interest.

Notes

1 For a detailed typology of discursive genres in English for Police Purposes, see Cartron (2022, pp. 173–96).

2 The term *probable cause affidavit* dominates, but it can vary depending on the police forces. Several designations have been identified: *affidavit of (or for) probable cause*, *affidavit for an arrest warrant*, *arrest affidavit*, *charging affidavit*, *complaint affidavit*, *probable cause affidavit*, *probable cause letter,* and *probable cause statement* (or *statement of probable cause*). Despite the variety of names used to designate this type of specialised text (affidavit, statement, or letter), their content and purpose remain identical.

3 *Affidavit* is a term borrowed from the medieval Latin *affidavit*, third person singular of the perfect indicative of *affidare*, which means "to declare under oath".

4 The presumption of innocence is based on the principle that a person is innocent until proven guilty.

5 The *Smoking Gun* website is famous for proving, in 2008, that an article in the *Los Angeles Times* entitled "An Attack on Tupac Shakur Launched a Hip-Hop War" was based on false documents, which led the newspaper to withdraw the article and publish an official apology (Rainey 2008).

6 The optical recognition software is available online at https://ocr.space (accessed on 8 February 2021).

7 To efficiently analyse the collected documents and be able to easily identify the sources of studied items, a file was created for each text, and a standardised naming system was elaborated. The files were named as follows: PC[for probable cause]_[US Postal Service code for the state, for instance, LA for Louisiana]_[Police force]_[Year]. To indicate the police force, abbreviations were used, such as *PD* for a *Police Department*, *SO* for a *Sheriff's Office*, or *FBI* for the *Federal Bureau of Investigation*.

8 Names, addresses, and personal details were redacted to follow the ethical guidelines and policy of the journal.

9 The Concordance plot tool of AntConc shows where a search word or expression is located in the texts. The length of the text is represented by the width of the blue bar, and each hit is indicated as a vertical line within the bar.

References

Baldwin, John. 1993. Police interview technique: Establishing truth or proof? *The British Journal of Criminology* 33: 325–52. [CrossRef]
Banks, David. 2016. Diachronic aspects of ESP. *ASp* 69: 97–112. [CrossRef]
Beacco, Jean-Claude. 2004. Trois perspectives linguistiques sur la notion de genre discursif. *Langages* 1: 109–19.

Benneworth, Kelly. 2009. Police Interviews with Suspected Paedophiles: A Discourse Analysis. *Discourse & Society* 20: 555–69.
Bhatia, Vijay K. 1993. *Analysing Genre: Language Use in Professional Settings*. London: Longman.
Bhatia, Vijay K. 2017. *Critical Genre Analysis: Investigating Interdiscursive Performance in Professional Practice*. New York: Routledge.
Bowers, Josh. 2014. Probable Cause, Constitutional Reasonableness, and the Unrecognized Point of a "Pointless Indignity". *Stanford Law Review* 66: 987–1050.
Brodeur, Jean-Paul, and Dominique Monjardet. 2003. *Connaître la Police. Grands Textes de la Recherche Anglo-Saxonne*. Paris: Les Cahiers de la Sécurité Intérieure.
Callahan, Mike. 2019. The Consequences of False Statements and Deliberate Omissions in Warrant Affidavits. *Lexipol—Police1*. Available online: https://www.police1.com/legal/articles/the-consequences-of-false-statements-and-deliberate-omissions-in-warrant-affidavits-6GQ5yBXnT7nitksb/ (accessed on 3 July 2022).
Carr, David. 2008. Dirty Job, but Someone Has to Do It. *The New York Times*. April 14. Available online: https://www.nytimes.com/2008/04/14/business/media/14carr.html (accessed on 12 June 2020).
Cartron, Audrey. 2022. Caractérisation de l'anglais de la police en tant que langue de spécialité: Contribution à l'élaboration d'un savoir savant visant à la construction d'un savoir à enseigner. Ph.D. thesis, Aix-Marseille Université, Aix-en-Provence, France.
Cartron, Audrey. 2023a. Présentation, description et enseignement d'un genre spécialisé: Les suspect interviews (auditions de mis en cause). In *Les Genres en Anglais de Spécialité: Définitions, Méthodologies D'analyse et Retombées Pédagogiques*. Edited by Margaux Coutherut and Gwen Le Cor. Berne: Peter Lang, pp. 137–65.
Cartron, Audrey. 2023b. A Study of the Psycho-Social Functions of Humour in English for Police Purposes. In *English for Specific Purposes and Humour*. Edited by Shaeda Isani and Michel Van der Yeught. Newcastle upon Tyne: Cambridge Scholars Publishing, pp. 63–90.
Charaudeau, Patrick. 2009. Dis-moi quel est ton corpus, je te dirai quelle est ta problématique. *Corpus* 8: 37–66. [CrossRef]
Charman, Sarah. 2013. Sharing a Laugh: The Role of Humour in Relationships Between Police Officers and Ambulance Staff. *International Journal of Sociology and Social Policy* 33: 152–66. [CrossRef]
Chiron, Pierre. 2007. *Aristote, Rhétorique (Présentation et Traduction par Pierre Chiron)*. Paris: Flammarion.
Coulthard, Malcolm. 2002. Whose Voice Is It? Invented and Concealed Dialogue in Written Records of Verbal Evidence Produced by the Police. In *Language in the Legal Process*. Edited by Janet Cotterill. Houndmills: Palgrave Macmillan, pp. 19–34.
Crespo, Andrew Manuel. 2020. Probable Cause Pluralism. *The Yale Law Journal* 129: 1276–391. [CrossRef]
Dando, Coral, Rachel Wilcock, and Rebecca Milne. 2009. The Cognitive Interview: Novice Police Officers' Witness/Victim Interviewing Practices. *Psychology, Crime & Law* 15: 679–96.
Fielding, Nigel. 1994. Cop canteen culture. In *Just Boys Doing Business? Men, Masculinities and Crime*. Edited by Tim Newburn and Elizabeth A. Stanko. London: Routledge, pp. 46–63.
Fox, Gwyneth. 1993. A Comparison of 'Policespeak' and 'Normalspeak': A Preliminary Study. In *Techniques of Description: Spoken and Written Discourse*. Edited by John Sinclair, Michael Hoey and Gwyneth Fox. London: Routledge, pp. 183–95.
Gaines, Philip. 2011. The Multifunctionality of Discourse Operator *Okay*: Evidence from a Police Interview. *Journal of Pragmatics* 43: 3291–315. [CrossRef]
Gayadeen, Shashi Marlon, and Scott W. Phillips. 2016. Donut Time: The Use of Humor Across the Police Work Environment. *Journal of Organizational Ethnography* 5: 44–59. [CrossRef]
Glaister, Dan. 2006. US Police Replace Codes With Plain English. 10–4? *The Guardian*. November 14. Available online: https://www.theguardian.com/world/2006/nov/14/usa.topstories3 (accessed on 21 December 2019).
Grabowicz, Paul. 2014. Tutorial: Police Records. Berkeley Graduate School of Journalism. Available online: https://multimedia.journalism.berkeley.edu/tutorials/police-records/ (accessed on 4 September 2019).
Hall, Philip. 2008. Policespeak. In *Dimensions of Forensic Linguistics*. Edited by John Gibbons and Teresa Turell. Amsterdam: John Benjamins, pp. 67–94.
Haworth, Kate. 2006. The Dynamics of Power and Resistance in Police Interview Discourse. *Discourse & Society* 17: 739–59.
Hepburn, Alexa. 2003. *An Introduction to Critical Social Psychology*. London: SAGE Publications.
Heydon, Georgina. 2013. From Legislation to the Courts: Providing Safe Passage for Legal Texts through the Challenges of a Police Interview. In *Legal-Lay Communication: Textual Travels in the Law*. Edited by Chris Heffer, Frances Rock and John Conley. Oxford: Oxford University Press, pp. 55–77.
Holdaway, Simon. 1988. Blue Jokes: Humour in Police Work. In *Humour in Society: Resistance and Control*. Edited by Chris Powell and George E. C. Paton. Houndmills: Macmillan Press, pp. 106–22.
Hyland, Ken. 2005. Stance and Engagement: A Model of Interaction in Academic Discourse. *Discourse Studies* 7: 173–92. [CrossRef]
Johnson, Edward. 2003. Talking Across Frontiers: Building Communication Between Emergency Services. In *New Borders for a Changing Europe: Cross-Border Cooperation and Governance*. Edited by O'Liam Dowd, James Anderson and Thomas M. Wilson. London: Frank Cass Publishers, pp. 89–111.
Johnson, Edward, Mark Garner, Steve Hick, and David Matthews. 1993. *PoliceSpeak: Police Communications and Language and the Channel Tunnel—Report*. Cambridge: PoliceSpeak Publications.
Kanoksilapatham, Budsaba. 2007. Introduction to move analysis. In *Discourse on the Move*. Edited by Douglas Biber, Ulla Connor and Thomas A. Upton. Amsterdam: John Benjamins Publishing Company, pp. 23–41.

Kinports, Kit. 2010. Veteran Police Officers and Three-Dollar Steaks: The Subjective/Objective Dimensions of Probable Cause and Reasonable Suspicion. *Journal of Constitutional Law* 12/3: 751–84.
Komter, Martha. 2001. La construction de la preuve dans un interrogatoire de police. *Droit et société* 48: 367–93. [CrossRef]
Leo, Richard A. 1996. Inside the Interrogation Room. *The Journal of Criminal Law and Criminology* 86: 266–303. [CrossRef]
Library of Congress. n.d. Constitution of the United States, Fourth Amendment. Available online: https://constitution.congress.gov/constitution/amendment-4/ (accessed on 13 October 2023).
Magid, Laurie. 2001. Deceptive Police Interrogation Practices: How Far Is Too Far? *Michigan Law Review* 99: 1168–210. [CrossRef]
Martin, Jacky. 1997. Du bon usage des corpus dans la recherche sur le discours spécifique. *ASp* 15–18: 75–83. [CrossRef]
Milne, Rebecca, and Ray Bull. 2006. Interviewing Victims of Crime, Including Children and People with Intellectual Disabilities. In *Practical Psychology for Forensic Investigations and Prosecutions*. Edited by Graham Davies and Mark R. Kebbell. Chichester: Wiley, pp. 7–23.
Nesi, Hilary. 2013. ESP and Corpus Studies. In *The Handbook of English for Specific Purposes*. Edited by Brian Paltridge and Sue Starfield. Malden: Wiley-Blackwell, pp. 407–26.
Oxburgh, Gavin, Trond Myklebust, and Tim Grant. 2010. The Question of Question Types in Police Interviews: A Review of the Literature From A Psychological and Linguistic Perspective. *Journal of Speech, Language and the Law* 17: 45–66. [CrossRef]
Petit, Michel. 2002. Éditorial. *ASp* 35–36: 1–2. [CrossRef]
Petit, Michel. 2010. Le discours spécialisé et le spécialisé du discours: Repères pour l'analyse du discours en anglais de spécialité. *E-rea* 8. [CrossRef]
Philbin, Tom. 1996. *Cop Speak: The Lingo of Law Enforcement and Crime*. New York: John Wiley & Sons.
Poteet, Lewis J., and Aaron C. Poteet. 2000. *Cop Talk: A Dictionary of Police Slang*. Lincoln: Writers Club Press.
Rainey, James. 2008. The Times Apologizes Over Article on Rapper. *Los Angeles Times*. March 27. Available online: https://www.latimes.com/local/la-me-tupac27mar27-story.html (accessed on 12 June 2020).
Reiner, Robert. 2000. *The Politics of the Police*. Oxford: Oxford University Press.
Richardson, Adam. 2018. PC for Writers, Wording for Warrants, Chain of Evidence for Murder Weapons—004. *Podcast*. August 17. Available online: https://www.writersdetective.com/pc-for-writers-wording-for-warrants-chain-of-evidence-for-murder-weapons-004/ (accessed on 10 March 2020).
Rock, Frances. 2001. The Genesis of a Witness Statement. *The International Journal of Speech, Language and the Law* 8: 44–72. [CrossRef]
Rock, Frances. 2007. *Communicating Rights: The Language of Arrest and Detention*. Basingstoke: Macmillan.
Rock, Frances. 2017. Recruiting Frontstage Entextualisation: Drafting, Artefactuality and Written-ness as Resources in Police-Witness Interviews. *Text and Talk* 37: 3–38. [CrossRef]
Rock, Frances. 2018. 'Apparently the Chap is a Bit of a Rogue': Upgrading Risk in Non-Emergency Telephone Calls to the Police. *Journal of Applied Linguistics and Professional Practice* 13: 4–37. [CrossRef]
Roulet, Eddy. n.d. Glossaire Français de Terminologie Linguistique. Analyse Modulaire du Discours: Définitions, Terminologie, Explications. Available online: https://feglossary.sil.org/sites/feglossary/files/amdfr.pdf?language=fr (accessed on 21 January 2021).
South Carolina Law Enforcement ETV Training Program. 1976a. Probable Cause for Arrest: Part I. Available online: https://www.ncjrs.gov/pdffiles1/Digitization/17400NCJRS.pdf (accessed on 21 November 2020).
South Carolina Law Enforcement ETV Training Program. 1976b. Probable Cause for Arrest: Part II. Available online: https://www.ncjrs.gov/pdffiles1/Digitization/17401NCJRS.pdf (accessed on 21 November 2020).
Stark, Jessica. 2020. A Contribution to the Characterisation of English for Diplomacy: Language, Discourse and Culture in the British Foreign and Commonwealth Office and the U.S. Department of State. Ph.D. thesis, Aix-Marseille Université, Aix-en-Provence, France.
Swales, John M. 1990. *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge University Press.
Taslitz, Andrew E. 2010. What is Probable Cause, and Why Should We Care?: The Costs, Benefits and Meaning of Individualized Suspicion. *Law and Contemporary Problems* 73: 145–210.
The Smoking Gun. 2020. About The Smoking Gun. Available online: http://www.thesmokinggun.com/about (accessed on 3 June 2020).
Tracy, Sarah J., and Karen Tracy. 1998. Rudeness at 911: Reconceptualizing Face and Face Attack. *Human Communication Research* 25: 225–51. [CrossRef]
Van der Yeught, Michel. 2016. Protocole de description des langues de spécialité. *Recherche et pratiques pédagogiques en langues de spécialité—Cahiers de l'APLIUT* 35: 1–71. [CrossRef]
Wozniak, Séverine. 2011. Contribution à la caractérisation de l'anglais de l'alpinisme, par l'étude du domaine spécialisé des guides de haute montagne états-uniens. Ph.D. thesis, Université Bordeaux 2, Bordeaux, France.
Wozniak, Séverine. 2019. *Approche Ethnographique des Langues Spécialisées Professionnelles*. Berne: Peter Lang.

languages

MDPI

Article

The Phraseology of Legal French and Legal Popularisation in France and Canada: A Corpus-Assisted Analysis

Manon Bouyé [1,†] and Christopher Gledhill [2,*,†]

[1] Equipe LEADS, Département d'Enseignement et de Recherche Langues, École Normale Supérieure Paris-Saclay, 91190 Gif-sur-Yvette, France; manon.bouye@ens-paris-saclay.fr
[2] Laboratoire CLILLAC-ARP, UFR EILA, Faculté Sociétés et Humanités, Université Paris Cité, 75006 Paris, France
* Correspondence: christopher.gledhill@u-paris.fr
† These authors contributed equally to this work.

Abstract: The popularisation of legal knowledge is a critical issue for equal access to law and justice. Legal discourse has been justly criticised for its obscure terminology and convoluted phrasing, which notably led to the Plain Language Movement in English-speaking countries. In Canada, the concept of Plain Language has been applied to French since the 1980s due to the official policy of bilingualism, while the concept has only been recently discussed in France. In this paper, we examine the impact of Plain Language rewriting on legal phraseology in French popularisation contexts. The first aim of our study is to see if plain texts published in France contain more traces of legal phraseology than French Canadian texts. Our second objective is to determine if a 'phraseology of plain language' can be identified across genres and languages. To do this, we compare two corpora of expert-to-expert legal texts written in French—made up, respectively, of legislative texts published in France and judicial texts published by the Supreme Court of Canada—with two corpora of texts that are claimed to have been written in Plain French Language for a non-expert readership—texts that guide laypersons through legal and administrative processes in France and summaries of decisions by the Supreme Court of Canada. Using n-grams, we extract and discuss the patterns that emerge from the corpora. In particular, our analyses rely on the concept of 'lexico–grammatical patterns', defined as the minimal unit of meaningful text made up of recurrent sequences of lexical and grammatical items. We then identify a sample of recurring lexico–grammatical patterns and their discursive functions.

Keywords: legal French; plain French; plain language; popularisation; phraseology

Citation: Bouyé, Manon, and Christopher Gledhill. 2024. The Phraseology of Legal French and Legal Popularisation in France and Canada: A Corpus-Assisted Analysis. *Languages* 9: 107. https://doi.org/10.3390/languages9030107

Academic Editors: Jeanine Treffers-Daller, Julien Longhi and Nadia Makouar

Received: 15 November 2023
Revised: 8 February 2024
Accepted: 26 February 2024
Published: 19 March 2024

1. Introduction

Plain Language (PL) is an attempt to encourage official institutions and other organisations to communicate with laypeople using accessible, clear, user-friendly language. The PL movement first gained momentum in the 1970s and 1980s in English-speaking countries: notably, in the United Kingdom, Australia, and New Zealand and then the United States and Canada (Asprey 2004). The initial domain of application for PL was legal and judicial contexts, such as the drafting of contracts and statutes, but the concept has since spread to other areas, such as public administration and medicine. PL differs from more formal language schemes (such as Basic English, Controlled Language, and so on) in that it corresponds to a nebulous series of stylistic preferences rather than an explicitly defined set of re-writing rules or vocabulary. The guidelines for PL include negative advice (avoid the passive, do not use rare or specialised terms, avoid complex verbs and complex prepositions, etc.) as well as more positive recommendations (use shorter sentences, prefer direct expressions, address the reader as 'you', etc.) (Cutts 2008; Williams 2004). Notwithstanding a lack of formal definition, PL has attained a high degree of official recognition in various English-speaking countries and has been implemented in both expert-to-expert communication (as in statutes) (Williams 2004, 2015) and expert-to-non-expert communication.

Many studies have been devoted to discussing the implementation of PL (Williams 2015, 2022) or its reception in various legal settings in English-speaking countries, including the United Kingdom and New Zealand (Masson and Waldron 1994; Rossetti et al. 2020). There has also been research from the point of view of discourse analysis that looks at popularisation and the dissemination of legal knowledge. In this paper, we focus on the popularisation of legal knowledge, defined by Engberg et al. (2018) as the recontextualisation of legal knowledge from contexts that exhibit a form of power asymmetry to a new non-expert context, with the intent of adapting the presentation of knowledge to the audience (Engberg et al. (2018) [our paraphrase]). As can be seen in this volume, numerous authors have investigated popularisation of the law carried out by legal institutions in various languages such as English (Cacchiani et al. 2018; Turnbull 2018), German (Luttermann, and Engberg 2018) and French (Preite 2016, 2018), while others have examined popularisation in non-institutional contexts through various media, including studies of YouTube videos used by expert lawyers (Cavalieri et al. 2018), children's books (Diani 2018) or teaching applications of TV shows in legal English classes (Dąbrowski 2017). Our paper follows on from research on popularisation produced by legal institutions, as it focuses on legal information texts published by the French government and Plain Language Summaries of judgements published by the Supreme Court of Canada. Both genres are intended for a non-specialised audience, and to the best of our knowledge, these genres have seldom been compared.

Generally speaking, the concept of PL is less well established outside the English-speaking world, and there have consequently been far fewer studies on PL or on the principles of clear language, especially in languages such as French. However, it is notable that the principles of PL have been widely adopted and implemented in Canada, and therefore also in the French-speaking parts of Canada, as part of bilingual language policy under the name *Langue Claire et Simple* (Simple and Clear Language). Various legal professions (barristers) and institutions (the Supreme Court of Canada) now claim to use PL in Canada (Asprey 2004). In contrast, the concept of PL is not as well-developed in France, and it has not achieved the same level of institutional recognition. One reason for this may be that relations between French citizens and the administration (*Service Public*) are notoriously difficult. This has been argued by both linguists (Collette et al. 2002) and independent observers, who have pointed out the complexity of legal and administrative procedures for French users of the law (des Droits 2019). Thus, although there have been attempts to implement the principles of PL in some contexts, it strikes us that there is generally still a considerable gap between the user-friendly discourse adopted by 'public-facing' organisations in many English-speaking contexts and the highly elaborate 'techno-heavy' style of French official discourse.

These general observations lead us to test two assumptions in this paper. In the first instance, we set out to test the hypothesis that French texts from France (FR FRA) are 'less simplified' than comparable texts from English-speaking countries (EN UK, EN NZ, etc.). This hypothesis was tested and partly confirmed in Bouyé (2022). In this paper, we test the related hypothesis that Canadian-French texts (FR CAN) are 'more simplified' than their European French counterparts (FR FRA). By 'more or less simplified', we are not talking about a single quantifiable characteristic, but rather, we are talking about two different configurations that can be identified systematically in two types of discourse (expert legal texts vs. plain legal texts). More specifically, abundant research has established that legal language is characterised by a 'highly nominal' style (Crystal and Davy 1969)—one of the characteristics that legal French shares with English, amongst other languages—along with Latin and Latinate terms, set formulae, a formal register, complex syntax due to a high degree of subordination, and the use of complex prepositional phrases (Galonnier 1997). One of our objectives is to examine the impact of the reconfiguration of legal knowledge on certain syntactic and stylistic features, including nominal and prepositional forms, as PL guidelines often encourage the use of verbal rather than nominal forms (Plain English Campaign 2022).

In the remaining sections of this paper, we explore these questions quantitatively and through the prism of phraseology. Several analysts have examined the phraseology of legal language, with a number of studies looking at regular expressions and lexical bundles in English (Biel 2017; Breeze 2017; Goźdź Roszkowski and Pontrandolfo 2013; Goźdź-Roszkowski and Pontrandolfo 2017). However, fewer studies have been conducted on the phraseology of PL in French. What we mean by 'the phraseology of plain language' is simply the typical wordings (routine formulae, extended collocations, lexico–grammatical patterns, etc.) that can be seen as statistically significant in one type of text (popularized texts, mediated knowledge, etc.) when compared to other types of text. More specifically, our concept of 'phraseology' corresponds to recurring sequences of lexico–grammatical sequences that operate as whole semantic units and serve a regular discourse function within a specific type of discourse or genre, a notion we have explored elsewhere (Bouyé and Gledhill 2019). Thus, in popularised legal texts, it is possible to identify recurrent sequences of this type (*This is called . . . Dans un délai de*) and to associate these sequences with specific discourse functions. These discourse functions include referential functions, i.e., sequences that refer to participants or elements of the legal process, as well as metatextual functions, such as sequences that define a term or direct the reader towards another part of the text or towards another text. In one of the first studies of this type (Bouyé and Gledhill 2019), we attempted to set out some characteristics of PL phraseology in English and French using n-grams. In this paper, we return to the concept of the 'lexico–grammatical pattern' (LxGr) in order to examine whether there is such a phenomenon as the 'phraseology of simplification'. In order to grasp the implications of this, it is important to provide a more formal account of LxGr patterns. We define LxGr patterns (Gledhill et al. 2017) as recurrent sequences of lexical items ('collocations') that correspond to regular grammatical structures[1] and that have a recognisable frame of reference or discourse function. Thus, each LxGr pattern corresponds to a 'minimal meaningful unit of text'. Unlike n-grams and other fixed sequences, LxGr patterns are productive and potentially discontinuous. The simplest forms of LxGr patterns are routine formulae or 'speech acts' (such as greetings, warnings, official pronouncements, etc.) (Gledhill et al. 2017).

In addition to our hypotheses regarding Plain Language in legal discourse, in this paper we also examine a number of more general research questions regarding LxGr patterns. In particular: what is the smallest possible sequence of items (n-gram) that corresponds to an LxGr pattern (i.e., a meaningful unit of text)? Furthermore, given a random selection of n-grams, is it possible to predict the discourse function for that LxGr (e.g., definition, procedure, explanation, evaluation, etc.)?

Returning more specifically to the topic of Plain Language in French, this paper considers the following research questions:

- Are there characteristic LxGr patterns in legal texts (i.e., in non-plain legal texts)?
- Are there traces of such patterns in PL texts?
- Similarly, are there characteristic LxGr patterns of PL in administrative discourse?
- More specifically, is it possible to establish a difference between generic phraseology (belonging to several 'genres') and specific phraseology (patterns that are 'unique' or at least more salient in one genre as opposed to all the others)?

We propose to answer these questions in the following sections. Section 2 introduces our data and corpus tools. Sections 3 and 4 present, analyse and discuss the results obtained from our data.

2. Materials and Methods

2.1. Data

The textual data used in this study are based on two French-language corpora: one consisting of popularisation texts destined for non-expert law users, entitled PLAIN, and the second, entitled LEX, made up of expert-to-expert written legal genres. Each of these corpora is subdivided into two subcorpora.

Concerning the PLAIN corpus, as mentioned above, this paper focuses on the popularisation of legal knowledge in French published by two government institutions. Turnbull (2018) distinguishes between 'popularisation', defined as the recontextualisation of information with the aim of broadening the reader's general knowledge, and 'knowledge mediation', in which information is transferred with the aim of allowing readers to take action performatively and thus to 'empower' themselves. The two popularisation genres represented here can be said to be instances of both mediation and popularisation as they relate directly to citizens' access to justice (*accès au droit et à la justice*) and public understanding of rights (*connaître et faire valoir ses droits*) and aim to help their readers make sense of the decisions taken by major judicial institutions or guide them through various legal and administrative processes. The first PLAIN subcorpus is composed of texts published by one of the most popular public service dissemination websites in France: *Service Public*. The texts are drafted and published by the governmental agency *Direction de l'information légale et administrative* (DILA), which is a department of the French Prime Minister's Office. It is one of the main public service legal mediators in the country. This subcorpus, entitled FR-Admin-DILA, is made up of 337 texts published between 2017 and 2019 and 466,472 word tokens. The second PLAIN subcorpus was collected from the Canadian Supreme Court website. It is made up of 66 summaries of decisions taken by the Canadian Supreme Court between 2017 and 2019. The small size of this corpus, which comprises 68,025 word tokens, can be explained by the specific type of text it is made of. The summaries, or 'Cases in Brief', are short summaries of the Court's judgements that recall the facts, explain the final decision reached by the Court and explain the positions of both the majority and dissenting opinions. This corpus is called FR-CA-Résumés.

As for the LEX corpus, it comprises FR-LAW, which contains articles and excerpts from statutes drafted between 1967 and 2018 and which were still in force at the time the corpus was compiled. It is representative of the legislative register and contains 5,083,750 word tokens. The second specialised corpus, entitled CA-Judgements, is made up of judicial texts: namely, decisions written and delivered by the Justices from the Supreme Court of Canada published between 2017 and 2018. It contains a total of 682,294 word tokens.

2.2. Methodology

Although in this paper we are focusing on phraseology, the first step in our analysis is to establish candidate sequences in the form of n-grams. In this case, we are interested in n-grams that identify parts of speech (POS) rather than just lexical forms. To identify salient POS-grams, we use the n-gram function of the concordance software Sketch Engine. Part-of-speech tags (not words) are used as attributes in order to extract not only major lexical bundles but also possibly salient syntactic regularities that could characterise our corpora. POS-grams allow us to capture the most salient grammatical constructions that may be candidate forms for more 'recognisable' LxGr patterns (as mentioned below, not all n-grams are potential LxGr patterns). To extract the most salient POS-grams in these corpora, a general reference corpus was used. This was the French Web Corpus (Sketch Engine) 2017, which is part of the TenTen family, a set of corpora obtained through web-crawling (Jakubíček et al. 2013) that contains a variety of text types, including news texts. We consider the French TenTen to be a general reference corpus. It must be noted that although this reference corpus contains more than four billion word tokens, only the first million word tokens in the corpus are used when computing the frequency on Sketch Engine. Key POS-gram candidates were identified based on Sketch Engine's keyness score feature, which uses the 'simple math method' (Kilgarriff 2009) to identify key words or key n-grams based on the normalised frequency of the word or n-gram in the focus corpus in relation to its normalised frequency in the reference corpus and includes a smoothing parameter. For each subcorpus, POS-grams with a keyness score above 100 were marked as LxGr-candidates. This means they were found to be at least a hundred times more frequent in the subcorpora than in the French TenTen reference corpus. To ensure that the selected patterns were over-represented in the legal or plain legal corpora as compared to

the reference corpus, chi-square tests were performed for each dataset using R. This point also applies to the other results we mention in Section 4 of this paper.

The n-gram function with POS tags as attributes returned a list of POS tag sequences that had to be converted back into readable POS patterns or lexical bundles. The analysis of concordances for each candidate POS-gram thus had to be carried out to identify LxGr patterns and their functions. Many POS-gram sequences with very high keyness scores corresponded to noise, i.e., they referred to numbers, prices or abbreviations in the various corpora. Others corresponded to parts of larger LxGr patterns. It was therefore necessary to consider POS-grams using contextual analysis and concordances. In Section 3, we present some POS-grams and patterns that were both highly recurring and interesting in terms of their rhetorical functions. This explains why some of our corpora have more distinctive patterns than others. As mentioned above, we do not attempt to make a simple distinction between 'specialised' and 'popularised' phraseology, but we are attempting to identify the typical patterns that emerge in comparable corpora, which can be characterised as expert-oriented ('specialised') and non-expert oriented ('popularised'). In the following discussion, we see examples of salient phraseology that can be identified as typical (in a statistical sense) in one corpus and atypical (or absent) in the other or can be observed as occurring in both corpora. This does not mean that we are in a position to fully characterise the phraseology of simplified language; rather, we claim here simply that it is possible to identify a representative sample of the most salient (outstanding, archetypical) phraseological units in our corpora, thus paving the way for a more complete analysis using other methods (e.g., textometrics).

We then performed a qualitative analysis of salient sequences based on domain-specific discourse functions put forward by Goźdź-Roszkowski et al. (2012). These include the category 'legal reference bundles': in particular, Institutional bundles, which refer to institutions; or Terminological bundles, which refer to specialised terms. The second category that was used to classify the POS-grams is text-oriented bundles: in particular, Structuring bundles. The final category used by Goźdź-Roszkowski et al. (2012) is Stance-oriented bundles, which contains Attitudinal bundles and Epistemic Stance bundles. Not all of these categories are represented in the results below.

In the rest of this paper, we use the following typographic conventions to refer to LxGr patterns. Lexical items are shown in italics, and each LxGr pattern is presented between angular brackets < >. All frequencies given in Tables are relative frequencies per million words (pmw).

3. Results

3.1. Analysis of Lexicogrammatical Patterns in Legal Texts

Table 1 shows some of the key POS-grams that were extracted from the FR-Law corpus from France.

Table 1. Key POS-grams in FR-Law Corpus (statutes).

Pattern	Relative Frequency in FR-Law Corpus	Relative Frequency in General Reference Corpus	Keyness Score
<Prep + *l'article* + Capital letter + Number [1]>	3151.6	4.8	540.8
<Prep + N masc + Prep + Det + N fem>	1545.1	64.3 1	23.6
<N + *prévues/fixées* + Prep + article + N>	795.3	2.5	231.8
<*le cas échéant*> [2]	623.36	1.76	

[1] Translation: 'Prep + section/article + capital letter + number'. [2] Translation: 'if/when applicable'.

Table 2 presents the key POS-grams in the corpus of Supreme Court judgements written in French.

Table 2. Key POS-grams in the CA-Judgements Corpus (Supreme Court judgements).

Pattern	Relative Frequency in CA-Judgements	Relative Frequency in General Reference Corpus	Keyness Score
<*La Cour d'appel* [1]>	648.2	0.4	461.5
<Prep + *l'article* + Num> [2]	965.3	0.011	955.6
<*Les juges majoritaires*> [3]	401.5	0.02	402.5 [3]
<Prep + det + N + prep + N>	612.6	3.7	129.9

[1] Translation: 'Prep + section/article + number'. [2] Translation: 'the Court of Appeal'. [3] Translation: 'the majority'.

Some patterns found in the results are specific (unique) to the type of discourse, or register, as can be seen when comparing the relative frequencies of these POS-grams in the two focus corpora, where they occur several hundred or even thousand times per million words, whereas they only appear a few times or a few dozen times in the general reference corpus. For example, Table 2 shows two specific 3-grams from the CA-Judgements corpus: *la Cour d'Appel*, which means 'the Court of Appeal' and *les juges majoritaires*, 'the majority judges'. Both are fragments of phrases that belong to longer Institutional bundles. These phrases refer to central figures and institutions of the Common law system: namely, the judges from the lower court along with the Court of Appeal, whose decision is the basis for a case being brought before the Supreme Court. In the case of the FR-Law corpus, one recurring pattern is *le cas échéant*, which means 'if the [aforementioned] conditions are fulfilled'. As we will see, these are key terms and phrases in legal discourse that can also be found in PLAIN texts but with only marginal recurrence (20 pmw).

What can be noted from Tables 1 and 2, however, is that both genres share common LxGr patterns, although with some variation. Most of these patterns correspond to what Goźdź-Roszkowski et al. (2012) calls Textual bundles and refer to other statutes or decisions or to other articles or sections in the same statute or judgement. These patterns are linked to cross-referencing: a broader characteristic of legal discourse (Tiersma 2000) that can be found in both genres (statute or Supreme Court opinion). In statutes, Bhatia (1994) calls these Referential provisions and explains that they point intertextually to other legislative texts or passages within the same text.

(1) <***Selon l'art. 662*** > *du Code criminel, la personne inculpée d'une infraction qui n'a été prouvée que partiellement peut être déclarée coupable d'une infraction moindre et incluse.* (CA-Judgements)
Section 662 of the Criminal Code provides that where a person is charged with one offence, but only a part of that offence is proved, he or she may be convicted of a lesser, included offence. (Official Supreme Court of Canada translation)

(2) *La durée de la période prévue* <***à l'article L. 434-9*** > *est fixée à trois ans.* (FR-Law)
The duration of the period is set to three years **under art. L. 434-9**. (Our translation)

Such patterns (using a prepositional phrase in French and a thematised phrase in English) are pervasive in this type of expert discourse (statutes, decisions), which requires constant references to previous decisions, statutes and legal instruments.

Another particularly interesting candidate is represented by the sequence <Prep + N + prep + N + prep>. This is a good LxGr candidate since it consists of a highly regular recurring grammatical structure but also allows for a large amount of lexical variation. A quick examination of some concordance results reveals that this pattern corresponds to a complex prepositional sequence, i.e., a prepositional phrase that introduces another noun phrase or prepositional phrase(s).

(3) *Les présents pourvois portent* <***sur l'étendue de l'obligation de communication du ministère public en ce qui a trait aux***> *registres d'entretien des alcootests.* (CA-Judgements)
These appeals deal **with the scope of the Crown's disclosure obligations with respect to** maintenance records of breathalyzer instruments. (Official translation by the Supreme Court of Canada)

The fact that this pattern is highly salient in our French legal corpora is significant. The use of complex noun chains and prepositional cascades is considered a typical feature of legal language, especially when it is in the form of complex prepositions in expressions such as *for the purpose of*[2] (Bhatia 1983; Biel 2017; Coulthard et al. 2016). Even more significantly, according to PL drafting guides in English and French, complex nominals and prepositional chains are among those features that should be avoided by legal writers because they purportedly contribute to the heavy nominal style (Crystal and Davy 1969) of legal language. In addition, the packing of information in nominal or prepositional cascades increases lexical density in texts (Halliday 1994). We discuss this pattern further in Section 3.3 since it is also found to some extent in PLAIN texts from our corpora.

It is also interesting to note at this point that many of the patterns we have identified above can be found within other patterns (either embedded or adjacent), as is evident in the example below, an excerpt from the French-Law corpus, in which two patterns from Table 1 can be identified:

(4) *Tout travailleur de nuit bénéficie d'un suivi individuel régulier* <***de son état de santé***> <***dans les conditions fixées à l'article L. 4624-1***>. (French-Law corpus) Any night worker has the right to regular health check-ups pursuant to the conditions set out (Our translation)

The fact that patterns occur alongside or within other patterns (thus, we use the term 'chains of pattern' or cascades) corroborates the idea that phraseological patterns constitute the building block of specialised discourse and here, in particular, of legislative or judicial discourse. We now turn to the main LxGr patterns in the PLAIN corpus.

3.2. Analysis of Lexicogrammatical Patterns in PLAIN French Texts

3.2.1. In the European French Admin Corpus

The key POS-grams in the French administrative corpus addressing law users correspond to very specific n-grams (or lexical bundles). Table 3 shows the most frequent patterns.

Table 3. Key LxGr patterns in French administrative texts (FR-PL-Admin).

Pattern	Frequency (pmw)	Freq in General French Corpus	Keyness Score
<*Vous devez/pouvez* [1] + V>	4032.4	14.3	263.3
<*dans un délai de* [2] + Numeral Adj + *jours/mois/semaines*>	568.1	0.9	297.7
<Prep + N + prep + N + prep>	508.1	3.6	111.3
<*Où s'adresser ?*> [3]	428.7	0.02	418.7
<*de quoi s'agit-il ?*> [4]	237.9	0.02	233.3

[1] Translation: 'You must/can'. [2] Translation: 'Within X days/months/weeks'. [3] Translation: 'Where can I get help?' [4] Translation: 'What does this mean?'

Although many of these items are very short, we suggest that several of these sequences are fragments of longer LxGr patterns, which themselves constitute recognisably meaningful units of text (according to our definition above). Many of these sequences correspond to text-oriented patterns, framed as direct questions, which have a cohesive, organisational function. These LxGr patterns either introduce the definition of a just-mentioned term, as in Examples 5 and 6 below, or explain who or what institution the law users can contact next in the legal process they are involved in. In Example 7, the question 'Who should I contact?' is immediately followed by a sentence that gives the phone number of the Police.

(5) **<De quoi s' agit-il ?>** Le sursis simple dispense la personne condamnée de l'exécution de la peine prononcée (peine de prison et/ou d'amende). (FR-Admin-PL)
What does this mean? Suspension with probation allows convicted individuals to avoid a particular sentence (imprisonment or fine) (Our translation).

(6) <**De quoi s'agit-il ?**> Le contrôle judiciaire est une mesure qui soumet la personne mise en cause dans une affaire pénale à une ou plusieurs obligations, dans l'attente de son procès. (FR-Admin-PL)
<**What does this mean?**> Bail is a measure that can be given, with one of more conditions, to a person accused of a criminal offence while they wait for trial. (Our translation)

(7) **<Où s' adresser ?>** Police secours – 17 Par téléphone Composez le 17 en cas d'urgence concernant un accident de la route, un trouble à l'ordre public ou une infraction pénale.
Where to get help Police services – By phone – Call 17 in case of an emergency, traffic accident, disturbing the peace or criminal offence. (Our translation) (FR-Admin-PL)

Taken as a whole, these LxGr patterns are part of what discourse analysts such as Turnbull (2018) have called 'conversational turns'. As such, they represent significant structuring devices in the dissemination of legal knowledge. Drafters of legal knowledge mediation texts imagine and anticipate the questions law users might have and use interrogative sentences to structure their texts when explaining the steps of the legal procedure or situations related to a legal right. In the case of *De quoi s'agit-il ?* (Example 5), we have an explicit signal that the text is going to provide the definition of a term. As we see later on, this structure is comparable to <il y a + N>, in which the authors interrupt their exposition to provide an explicit metadiscoursal definition. The difference here is that <De quoi s'agit-il?> and <Où s'adresser> both belong to oral discourse and are explicit markers of turn-taking; whereas <Il y a X quand / lorsque> belongs to elaborate, expository discourse.

Other patterns correspond to lexical phrases that are are linked to the steps of the legal or administrative process itself, such as the complex prepositional phrase *dans un délai de + N*, which defines the time limit for legal action for law users. This highly frequent bundle, which appears only 0.9 pmw in the general French corpus, appears to be quite specific to the FR-Admin corpus. It is usually part of an extended LxGr pattern that contains an Actor (usually second-person *You*), a verb phrase referring to some type of legal action (such as reporting a crime or referring to an institution) and the complex prepositional phrase, which functions as a time adjunct.

(8) *À savoir : en raison des règles de prescription, vous devez déposer votre plainte pour viol* <***dans un délai de 20 ans***> *à compter de la date des faits.* (FR-PL-Admin)
Please be aware that because of the statute of limitation, you need to file a rape complaint **within twenty years**. (Our translation)

(9) *Si votre demande est acceptée, vous en êtes informé par courrier* <***dans un délai de 4 mois***>*.* (FR-PL-Admin)
If your request is approved, you will be notified **within four months**.

In Example 8, the time limit that is defined is twenty years, and the legal action is reporting a specific felony (rape). Interestingly, this type of pattern can also be found in the FR-Law corpus, as in the example below. Although the pattern <*dans un délai de*> appears in the legislative corpus, it is about half as frequent in the specialised corpus (229.5 pmw vs. 568.1 pmw in the PLAIN FR-Admin corpus). It thus seems to be relatively more specific to the discourse of administrative French.

(10) *L'autorité administrative statue sur la demande* <***dans un délai de six mois***> *à compter du dépôt par l'étranger du dossier complet de cette demande.* (FR-Law)
The authority reviews the application **within six months** after the application has been made. (Our translation)

Whereas the pattern is found in sequences wherein the reader is directly addressed in the PL administrative texts, the legislative excerpt (Example 10) is much more impersonal, as the subject of the verb is an abstract entity (*the administrative authority*).

In the specialised texts (FR-LEX, CA-Judgements), we find a number of sequences in which the subject of the verb corresponds very often to an abstract legal or administrative concept. In the administrative corpus, the typical subject of certain patterns corresponds

more often to the law user (thus representing a significant re-orientation of the discourse). This difference suggests what has been called a process of 'personalisation' (Turnbull 2018), by which expert legal knowledge is reformulated for a non-expert readership.

For example, one of the most common POS-grams we find in the corpus is <*Vous devez/pouvez* + V>, involving the second person *vous* and a modal verb expressing either obligation (*devez* = must) or possibility (*pouvez* = can/may). The lexical verb that is introduced in these contexts often expresses an administrative procedure (expressed as a Material or Behavioural process[3]).

(11) <***Vous devez** écrire*> *directement au procureur de la République.*
You must write to the public prosecutor directly (Our translation). (FR-Admin-PL)

(12) <***Vous pouvez** collecter*> *vous-même les preuves de ce harcèlement.*
You can collect evidence of your harassment yourself (Our translation). (FR-Admin-PL)

Similarly to *dans un délai de*, this pattern is usually associated with various types of legal action, albeit not as fixed in terms of the syntactic frame. The modals of both the obligation and possibility vary according to the way in which legal dissemination texts define the user's legal rights and obligations. As with previous examples of reorientation, the use of the second-person pronoun is part of the communication strategy called '*conversationalization* of public discourse' (Turnbull 2018), which is performed through the use of direct questions (as seen above) as well as first- or second-person pronouns to reformulate highly abstract legal knowledge and to create a form of dialogue between the institution and the non-expert readership.

3.2.2. Summary of the Canadian Plain Language Corpus

The most salient POS-grams and lexical bundles in the summaries of judgements are presented in Table 4.

Table 4. Main LxGr patterns in French Canadian summaries.

Pattern	Frequency (pmw)	Freq in General French Corpus	Keyness Score
<*La Cour d'Appel*> [1]	1161.2	0.4	826.2
<*La Cour Suprême a* + V + que> [2]	450.8	1.5	183.4
<*le droit/pouvoir de* + V	450.8	0.3	345.4
<*Les juges majoritaires ont* + V + *que* [3]>	778.7	0.16	674.4
<la Charte canadienne des droits et [libertés]> [4]	220.5	0.02	2061.3

[1] Translation: <The Court of Appeal>. [2] Translation: <The Supreme Court + V + that>. [3] Translation: <The majority + V + that>. [4] Translation: <The Canadian Charter of Rights and Freedoms>

The LxGr patterns in the CA-PL-Summaries are mostly Institutional bundles, referring to either institutions (the Court of Appeal), legal actors (judges) or to foundational legal texts: in particular, the Canadian Charter of Rights and Freedoms, which sets out and protects a number of rights and freedoms, as can be seen below.

(13) ***<La Cour d'appel>** a dit partager l'opinion du premier juge. (CA-PL-Summaries)*
The Court of Appeal agreed with the trial judge. (Official English version)

(14) *M. Chhina a fait valoir que son traitement était illégal au regard de <**la Charte canadienne des droits et libertés**>, qui fait partie de la Constitution du Canada.*
Mr. Chhina said that his treatment was illegal under the **Canadian Charter of Rights and Freedoms**, part of Canada's constitution. (Official English version of the summary by the Supreme Court of Canada)

It can be noted that several of the shorter LxGr patterns presented in Table 4 and found in the CA-PL-Summaries are the same as those found in the CA-Judgements corpus: namely, *la Cour d'appel* ('the Court of Appeal') or an extended LxGr pattern of the type *les juges majoritaires ont + V + que* ('the majority judges + V + that'). These mostly correspond

to key terms of judicial discourse in the Common Law system, which explains why they are used abundantly by the Supreme Court Justices in their judgements. As for the summaries, they have both an encapsulating and explanatory function. As such, they need to explain what the various courts involved in a case have decided and, in particular, what the majority opinion of the judges was; hence, there are frequent references to Courts of Appeal and to the majority judges.

What is interesting is that many of the sequences found through the analysis of POS-grams reveal extended LxGr patterns, often corresponding to projection (expression of engagement through indirect speech).

(15) <***Les juges majoritaires ont affirmé que***> *la police a porté atteinte aux droits que la Charte garantit à M. Reeves en prenant l'ordinateur sans son consentement et sans mandat. (CA-PL-Summaries)*
The majority said that the police breached Mr. Reeves' Charter rights by taking the computer without his consent and without a warrant. (English version of the summary by the Supreme Court of Canada)

(16) ***<La Cour suprême a confirmé que>****, dans une cause criminelle, le doute «raisonnable» doit être fondé sur la preuve et non sur des conjectures.*
In a criminal case, 'reasonable' doubt should be based on evidence, not speculation, **the Supreme Court has confirmed**. (English version of the summary by the Supreme Court of Canada)

These examples appear to belong to an extended LxGr pattern for which we suggest the following formula: <Institutional subject + Communicative process (state, confirm) + *que* (that) + Reporting clause>. This pattern is linked to a key inherent function of the summary genre, which is to report the Supreme Court's decisions by explaining the opinions expressed not only by the Justices but also by the inferior courts. Some authors discuss this in terms of 'discursive heterogeneity' (Preite 2016), referring to the fact that legal popularisation discourse (like all popularisation discourse) is based on a 'primary' specialised discourse that is explicitly mentioned as a legitimising source. This is especially true in the summaries of decisions, as the figure of the judge is at the core of the explanation and elaboration strategies.

3.3. Focus on Phraseology: Patterns from Legal Texts Also Found in the PLAIN Corpora

We have seen a certain number of patterns that appear to be specific to certain genres or are found in a certain type of register (legal or popularised text). We now turn to some specific patterns that seem to be of particular interest because they are particularly representative of legalese, and they can also be found in PLAIN legal texts from our corpus.

3.3.1. *Il y a* (There Is/There Are)

The first pattern we want to focus on is the existential or presentational structure *Il y a*, which is usually translated as *There is/are*. Although its keyness score is not extremely high, the pattern caught the authors' attention when analysing the data, as it is used with a specific syntax and is associated with an extended LxGr pattern that has a specific discourse function in the concordances in both legal and plain texts. It furthermore exhibits a higher relative frequency across our corpora as compared to the reference corpus (respectively: FR-Law 125.9 pmw; FR-Judgements 109.3 pmw; FR-Admin-PL: 132.9 pmw vs. FrenchTenTen 34.11 pmw). Some examples of this pattern and its use are set out in Figures 1 and 2.

LxGr spécifique aux corpus CA JUG, FR LOI: 'Defining (the conditions of) a Legal Term.'

< [Theme] ***Il y a*** + N (dénomination d'un délit / crime) ***lorsque*** + [Définition] >

A1) Est qualifié crime ou délit flagrant le crime ou le délit qui se commet actuellement, ou qui vient de se commettre. **Il y a** aussi **crime ou délit flagrant lorsque**, dans un temps très voisin de l'action, la personne soupçonnée est poursuivie par la clameur publique, ou est trouvée en possession d'objets …

A2) Les sanctions qui ne sont pas incompatibles peuvent être cumulées ; des dommages et intérêts peuvent toujours s'y ajouter. **Il y a force majeure** en matière contractuelle **lorsqu'**un événement échappant au contrôle du débiteur, qui ne pouvait être raisonnablement prévu lors de la conclusion du contrat…

A3) Annulation de l'acte instrumentaire162 **Il y a lieu à annulation lorsque** l'acte est irrégulièrement dressé, bien que ses énonciations soient exactes.

A4) … tout mur servant de séparation entre bâtiments jusqu'à l'héberge, ou entre cours et jardins, et même entre enclos dans les champs, est présumé mitoyen s'il n'y a titre ou marque du contraire. **Il y a marque de non-mitoyenneté lorsque** la sommité du mur est droite et à plomb de son parement d'un côté, et présente de l'autre un plan incliné.

A5) La violence est une cause de nullité qu'elle ait été exercée par une partie ou par un tiers. **Il y a** également **violence lorsqu'**une partie, abusant de l'état de dépendance dans lequel se trouve son cocontractant à son égard, obtient de lui un engagement qu'il n'aurait pas souscrit en l'absence d'une telle contrainte …

Figure 1. Concordances of presentational structure <*Il y a* + NG (term)> in French specialized legal texts.

< [Theme] ***Il y a*** + N (dénomination d'un délit / crime) ***si / quand*** + [Définition] >

B1) Cas d'abus de confiance. **Il y a abus de confiance quand** une personne s'approprie un bien que lui a confié sa victime. </s><s> Ce bien peut être une somme d'argent

B2) Différence avec vol et escroquerie. L'abus de confiance se distingue de l'escroquerie. **Il y a escroquerie si** l'auteur fait croire qu'il possède un droit sur le bien (par exemple, si l'auteur des faits retire de l'argent sur le compte de la victime avec une fausse procuration)

B3) Téléservice Harcèlement scolaire - Violences scolaires - Provocation au suicide. **Il y a harcèlement scolaire quand** un élève fait subir à un autre, de manière répétée, des propos ou des comportements agressifs.

B4) La simple tentative de vol ou de racket suffit pour rendre une plainte recevable. </s><s> **Il y a tentative si** l'auteur des faits a commencé à commettre son infraction mais qu'elle a échoué à cause d'un élément indépendant de sa volonté.

B5) Les violences conjugales peuvent correspondre à des violences : psychologiques, physiques, sexuelles, ou économiques. </s><s>. **Il y a violence conjugale quand** la victime et l'auteur sont dans une relation sentimentale.

Figure 2. Concordances of presentational structure <*Il y a* + NG (term)> in French PLAIN texts.

In the excerpts from FR-Law and CA-Judgements shown in Figure 1, this pattern is used to define a legal term or the conditions for a legal term to be established, especially a crime or felony. Similarly, in the excerpts from FR-Law and CA-Judgements shown in Figure 2, a similar pattern is used to specify and define a legal term. In the figures, the existential structure "*il y a*" is shown in black, while the term is presented red and the conjunction introducing a conditional clause or definition is in purple font. Significantly, this is always a term that has just been mentioned in the immediate co-text. Grammatically speaking, the indefinite article is always elided in this construction (before the term), which

is unusual in French: hence, our choice to our focus on this structure. This particular feature allows us to draw up a formula for the LxGr pattern as a whole: <[term in preceding discourse] ... *il y a* + [Ø] + [name of the term] // conditional clause + definition>. Finally, a notable difference in the corpora is that in the PL texts, the clause that introduces the definition or conditions for the crime or felony is *quand* (when) or *si* (if), while the legal texts contain its more formal alternative *lorsque*.

The structure *Il y a + [Ø article] + definition* happens to be used in other popularisation/dissemination discourses: for example, medical popularisation. The example below is an excerpt from the French National Health Service website explaining the signs of cardiac arrest, and it introduces the term using the same <*il y a* + [Ø]> construction.

(17) <***Il y a** arrêt cardiaque si> : la victime perd connaissance, tombe et ne réagit pas quand on lui parle ou qu'on la stimule*
It is a case of cardiac arrest if the victim is unconscious, falls or shows no reaction when talked to or stimulated (our translation).

Other types of dissemination discourse that use this pattern include promotional discourse or game rules. It appears to be specific to elaboration strategies and definitions of terms in French.

3.3.2. Prepositional Cascades

The second pattern we now turn to is often found in specialised legal texts and is especially highly frequent in the FR-Law corpus, although the CA-Judgements corpus also contains some examples as well as variations on the same type of structure. The pattern in question can be seen as an expanded version of the complex prepositional chains mentioned in Section 3.1, as it is composed of a chain of three prepositional phrases. What is surprising is that this highly complex structure can also be found routinely in the texts from the PLAIN corpus. Table 5 displays the frequency (pmw) of this POS-gram in the specialised corpora and their simplified versions as well as an example from each corpus.

Table 5. A salient POS-gram across corpora: Prepositional cascades <Prep + N + prep + N + prep>.

Corpus	Relative Frequency (pmw)	Example
FR-LAW	7150	Sont exonérés <**de droits de mutation par**> décès le conjoint survivant et le partenaire lié au défunt par un pacte civil de solidarité.
FR-Admin-PL	5702	Faire un recours préalable auprès de la MSA <**par courrier de préférence en**> recommandé avec avis de réception
CA-Judgements	2015	Les juges ordonnent fréquemment à des individus, <**à titre de condition à leur mise en liberté sous caution**>, d'éviter tout contact avec l'alcool et les drogues.
CA-Summaries	1502	Dans un tel cas, l'accusé peut interjeter appel de la déclaration <**de culpabilité pour meurtre au** > deuxième degré

What stands out in this table is that, first, the prepositional cascades <Prep + N + prep + N + prep> are particularly prominent not only in the FR-LAW corpus, with a frequency of 7150 pmw, but also in the FR-Admin-PL, with 5702 pmw. Although this is half as frequent in the CA-PL-summaries than in the CA-Judgements corpus, it is also statistically salient in the judicial dissemination corpus. This result suggests that the syntactic complexity that characterises legal language (Bhatia 1983), even after simplification to address a lay audience, still percolates into the plain texts under study. We discuss our results in the following section.

4. Discussion

The results of our phraseological exploration of legal French are consistent with other phraseological studies of legal language in other languages and contexts, as they underline the highly intertextual and syntactically complex nature of legal French in both

subcorpora. Our results also suggest differences in the two genres represented here, as the use of complex prepositions and prepositional cascades appears to be more frequent in the legislative FR-Law than in judicial discourse, represented by CA-Judgements, a result which is also consistent with previous research on legal phraseology in other languages, such as Spanish (Pontrandolfo 2021). Of course, the fact that the two legal subcorpora come from different countries, France and Canada, might also account for this difference in the frequency of these prepositional structures, as legal phrasemes are 'bound to a particular legal system' and the results should therefore be interpreted with caution (Pontrandolfo 2023). Future research on comparable judicial and legislative corpora in both European-French and French-Canadian would be necessary to examine whether the difference is genre-based or culture-based.

Concerning the PLAIN corpus, our preliminary results show that there are traces of legal phraseology in both dissemination corpora (that is to say, in the Canadian French corpus and the European French corpus). This becomes particularly clear in the case of projecting/reporting structures and complex prepositional chains. This leads us to suggest that the simplification of legal phraseology in French, though evident in both PLAIN corpora, is only achieved to a limited extent: an observation that is consistent with previous findings (Rossetti et al. 2020). There are, however, some LxGr patterns that appear to be characteristic of plain legal French, especially for certain genres. We note the very high frequency of lexical bundles like *De quoi s'agit-il?* ('What is it?') or *Où s'adresser* in the FR-Admin-PL corpus, as illustrated again by the example below in which both patterns create a dialogical structure in a text that is explaining what to do in case of harassment.

(18) <***De quoi s' agit-il ?***> *Le harcèlement est le fait de tenir des propos ou d'avoir des comportements répétés ayant pour but ou effet une dégradation des conditions de vie de la victime. (. . .)* <***Où s'adresser ?***> *La victime peut porter plainte contre le ou les auteurs du harcèlement. (FR-Admin-PL)*
What is it? Harassment is defined as repeated words of behaviours that aim at or result in the deterioration of the victim's living conditions. (. . .) **Where to get help** The victim can complain about the person(s) that are harassing them.

In our preceding analysis, we have seen lexical patterns that either serve to 'structure' the text (Halliday's textual metafunction), to introduce specialised terms (the *'Il y a'* construction) or draw the law users' attention to specific points in the text. Thus, we can characterise the overall strategy in the FR-DILA corpus as an attempt to provide readers with heuristic guidance—helping them to navigate around administrative procedures rather than setting out a 'boiled down' or simplified version of these procedures.

In the PL summaries from the Supreme Court of Canada, the communicative strategy is quite different: these texts are structured by references to the central figure—the judge—and the judicial institutions, which include the Supreme Court as well as the lower courts of appeal. Structurally, the core phraseology of this genre is oriented towards reporting clauses or to expansions (explanations), both of which involve extensive use of subordinate and embedded clauses. This can be seen in the example below.

(19) Selon **les juges majoritaires**, cela surcharge le système judiciaire,**qui** consacre plus d'argent à essayer de faire en sorte que des gens pauvres paient leur suramende qu'il en obtiendrait de ceux-ci. <**Les juges majoritaires ont**> fait observer <**qu**>'une peine est plus efficace si elle est adaptée à la personne.
The majority said this also burdened the justice system,**which** spent more trying to get poor people to pay the surcharge than it would ever get back. **The majority noted that** a sentence works best**if** it is made for the individual. (Official English version)

One of the notable differences between the two PLAIN corpora is their types of syntactic complexity. Our preliminary analysis suggests that the Fr-PL-Admin corpus is characterised by more nominal complexity, i.e., by the use of prepositional cascades and complex noun groups, while the CA-summary texts appear to contain more clausal complexity, i.e., more subordinate clauses, especially reported speech. In this sense, the

French administrative texts are closer to the legislative texts they are based on and are more complex in terms of information packaging in noun and prepositional phrases. The French Canadian corpus also exhibits more clausal complexity and is thus more 'elaborate' in this particular sense of clausal complexity, as can be seen in Example 19, in which we have emphasised relative pronouns as well as binding and linking conjunctions. Generally speaking, it might be expected that 'simplified discourse' will, in fact, turn out to be more elaborate (i.e., involve more structural expansions) than the highly codified but also much denser and compact 'expert discourse'. Indeed, greater clausal and verbal complexity linked to elaboration strategies and unpacking of information has also been shown in other PL discourse: in particular, medical discourse (Gledhill et al. 2019). The results we have set out above may also be due to the influence of the Plain Language Movement in French-speaking Canada, whereas the concept has not been institutionalised as such in France. If the FR-Admin corpus is closer to the FR-Law corpus, it may also be due to the fact that some of the texts published are actually direct quotations, without any simplification, of the original legal text, as suggested by results put forward by Bouyé (2022). Despite these differences in communicative strategies and complexity, our findings regarding 'plain French' appear to be consistent with recent research on legal popularisation; as is the case in other online lay-oriented discourse, drafters 'balance impersonal explanatory strategies with more interpersonal and communicative strategies' (Diani et al. 2023).

In the first section of this paper, we raised several questions about the nature and distribution of lexico–grammatical (LxGr) patterns. Our first question related to the relative 'size' of n-grams (What is the smallest sequence possible or useful to identify meaningful LxGr patterns?). In the data analysis above, we have demonstrated that it is often possible to analyse short n-grams as longer stretches of meaningful text that often correspond to specific discourse functions. The primary example here is the very short sequence *Il y a*, which on its own can only be seen as a short verbal group in French. Out of context, this sequence is so ubiquitous in French that it tells us very little, but once we begin to look at the corpus data, *Il y a* turns out to be a very distinctive definitional routine in both of the main FR corpora we analysed here.

Regarding our more general questions, it is useful to deal with each of them in turn:

(1) Is it possible to assign a discourse function to random n-grams, assuming that these sequences have been found to be salient in one of the subcorpora? Here, we have demonstrated this as a principle, although we have clearly not shown this positively across a wide range of data.

(2) Is it possible to identify characteristic LxGr patterns in legal texts (i.e., in non-plain legal texts)? This has been demonstrated using various examples, such as complex nominal groups/prepositonal phrases (Section 3.1).

(3) Is it possible to identify characteristic LxGr patterns of PL in administrative discourse? This has also been demonstrated in relation to turn-taking sequences and procedural constructions, among other examples (Section 3.2).

(4) Is it possible to establish a difference between generic phraseology (belonging to several 'genres') and specific phraseology (patterns that are 'unique' or at least more salient in one genre as opposed to all the others)? We have shown that certain constructions (such as the projection structures of reported speech) are 'generic' and occur as significant LxGr patterns in all of the corpora we have analysed here. This is especially evident in the two specialised corpora, in which LxGr patterns related to cross-referencing pervade both judicial and legislative corpora. Despite obvious macro-textual differences, the FR-Law and CA-Judgements corpora can be identified as belonging to the legal register as a whole; this is, notably, based on the 'generic' patterns related to cross-referencing, one of the features that gives legal language its characteristic 'legal flavour' (Maley 1994). Regarding specific LxGr patterns, we have identified a significant sample of these, among the most recognisable ones being the multiple prepositional phrase pattern (associated with FR Law and CA-Judgements)

as well as the *Il y a* definitional pattern (specific to the LEX corpus by use of the conjunction *lorsque* or to the PLAIN corpus by use of the subordinators *quand/si*).

5. Conclusions

In this paper, we have tried to characterise the phraseology of two legal genres and of two types of legal popularisation texts in French from France and from Canada. Our results suggest that salient n-grams and LxGr patterns can be identified in both legal French and PL discourse. Some patterns appear to be generic patterns, i.e., they are linked to the legal register as a whole, while others are specific to a genre or text type. In particular, our findings contribute to the characterisation of administrative and legal dissemination in French as phraseological patterns linked to knowledge recontextualisation and elaboration. Traces of highly complex phraseological patterns from legal language can be found in both PLAIN corpora, although they are more salient in European French texts, suggesting that these are more complex (in the sense of 'lexico–grammatically elaborate') than Canadian French texts. There are probably deep-rooted cultural reasons for this (higher expectations placed on French-speaking users of the law/public services, the relatively recent emergence of the plain language movement in France, differences in legal culture as well as practice . . .).

Further research is needed to confirm the preliminary observations set out in this article. Possible research perspectives include designing a survey to measure the comprehensibility of plain legal French by obtaining readability measures (self-paced reading or eye-tracking) from ad hoc tasks targetting specific features in legal texts: for example, complex prepositional phrases or passives.

Author Contributions: Conceptualisation, M.B. and C.G.; methodology and analysis C.G.; validation, M.B. and C.G.; formal analysis, M.B. and C.G.; investigation, M.B. and C.G.; resources, M.B.; data curation, M.B. and C.G.; writing—original draft preparation, M.B. and C.G.; writing—review and editing, C.G.; supervision, C.G.; project administration, M.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Publicly available datasets were analysed in this study. These data can be found here: https://www.scc-csc.ca/case-dossier/cb/index-fra.aspx accessed on 2 November 2023; https://www.service-public.fr/ accessed on 2 November 2023; https://www.legifrance.gouv.fr/ accessed on 2 November 2023.

Acknowledgments: The authors would like to thank the anonymous reviewers and coordinators of this issue.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

det	Determiner
LxGr	Lexico–Grammatical Pattern
N	Noun
NP	Noun Phrase
Num	Number
PL	Plain Language
pmw	Per Million Words
Prep	Preposition

Notes

1 i.e., 'constructions' in Goldberg's 1985 sense of the term, such as the passive, the resultative, etc.

2 Complex prepositions involve the embedding of one prepositional phrase within another (such as *for the purpose of this section*). Sometimes, these structures undergo lexicalisation in which the first prepositional group functions as a single preposition (Biel et al. 2015).

3 The terms in capital letters relating to Transitivity in the Systemic Functional Linguistics model are borrowed from (Halliday and Matthiessen 2013).

References

Asprey, Michele M. 2004. Plain Language around the world. In *Plain Language for Lawyers*, 3rd ed. Sydney: Federation Press.

Bhatia, Vijay K. 1983. Simplification v. Easification—The Case of Legal Texts. *Applied Linguistics* 4: 42–54. [CrossRef]

Bhatia, Vijay K. 1994. *Cognitive Structuring in Legislative Provisions*. London: Longman, vol. 1.

Biel, Lucja. 2015. Phraseological profiles of legislative genres: Complex prepositions as a special case of legal phrasemes in eu law and national law. *Fachsprache* 37: 139–160. [CrossRef]

Biel, Lucja. 2017. Lexical bundles in EU law: The impact of translation process on the patterning of legal language. In *Phraseology in Legal and Institutional Settings. A Corpus-Based Interdisciplinary Perspective*. London and New York: Routledge, vol. 10, p. 26.

Bouyé, Manon. 2022. Le style clair en droit: Étude comparative du discours juridique en anglais et en français, avant et après simplification. Le plain language dans la communication juridique avec le grand public. Ph.D. Thesis, Université Paris-Cité, Paris, France.

Bouyé, Manon, and Christopher Gledhill. 2019. Disseminating legal language for the general public: A corpus-based study of the discursive strategies used in english and french. In *Langues et Langages Juridiques. Traduction et Traductologie, Didactique et Pédagogie. Colloque International de Bordeaux*. Paris: Institut Francophone pour la Justice et la Démocratie, pp. 349–69.

Breeze, Ruth. 2017. Giving voice to the law: Speech act verbs in legal academic writing. In *Phraseology in Legal and Institutional Settings*. London: Routledge, pp. 221–39.

Cavalieri, Silvia. 2018. Broadcasting legal discourse. the popularization of family law through youtube. In *Popularization and Knowledge Mediation in the Law/Popularisierung und Wissensvermittlung im Recht*. Münster: LIT Verlag, pp. 251–70.

Cacchiani, Silvia. 2018. The voice of the law on gov.uk and justice.gouv.fr: Good value to citizens and institutions? In *Popularization and Knowledge Mediation in the Law*. Edited by Jan Engberg, Silvia Cacchiani, Karin Luttermann and Chiara Preite. Münster: LIT Verlag, pp. 117–48.

Collette, K., M. P. Benoît Barnet, D. Laporte, F. Pouëch, and B. Rui-Souchon. 2002. *Guide Pratique de la Rédaction Administrative*. Paris: Ministère de la Fonction Publique.

Coulthard, Malcolm, Alison Johnson, and David Wright. 2016. *An Introduction to Forensic Linguistics: Language in Evidence*. London: Routledge.

Crystal, David, and D. Davy. 1969. *Investigating English Style: English Language Series*. London and Harlow: Longmans.

Cutts, Martin. 2008. *Plain English Lexicon*. London: Plain Language Commission.

Dąbrowski, Andrzej. 2017. Reel justice in the context of teaching legal english as a foreign language. *Glottodidactica* 43: 107–20. [CrossRef]

des Droits, Défenseur. 2019. *Dématérialisation et Inégalités D'accès aux Services Publics*. Rapport. Paris: Défenseur des Droits.

Diani, Giuliana. 2018. Popularization of legal knowledge in English and Italian information books for children. In *Popularization and Knowledge Mediation in the Law.* Münster: Lit Verlag, pp. 291–316.

Diani, Giuliana. 2023. Disseminating legal information on online law forums in english and italian. *Ibérica* 46: 299–320. [CrossRef]

Engberg, Jan, Karin Luttermann, and Silvia Cacchiani. 2018. *Popularization and Knowledge Mediation in the Law/Popularisierung und Wissensvermittlung im Recht* Münster: LIT Verlag, vol. 9.

Galonnier, Bernard. 1997. Le discours juridique en France et en Angleterre. Convergences et spécificités. *ASp la Revue du GERAS* 15–18: 427–38. [CrossRef]

Gledhill, Christopher, Hanna Martikainen, Alexandra Mestivier, and Maria Zimina-Poirot. 2019. Towards a linguistic definition of 'simplified medical English': Applying textometric analysis to cochrane medical abstracts and their plain language versions. *LCM-La Collana/The Series* 91–114.

Gledhill, Christopher, Stéphane Patin, and Maria Zimina. 2017. Lexico-grammaire et textométrie: Identification et visualisation de schémas lexico-grammaticaux caractéristiques dans deux corpus juridiques comparables en français. *Corpus* 17. [CrossRef]

Goźdź-Roszkowski, Stanisław. 2012. Discovering patterns and meanings: Corpus perspectives on phraseology in legal discourse. *Roczniki Humanistyczne* 60: 47–70.

Goźdź Roszkowski, Stanisław, and Gianluca Pontrandolfo. 2013. Evaluative patterns in judicial discourse: A corpus-based phraseological perspective on american and italian criminal judgments. *International Journal of Law, Language and Discourse* 3: 9–69.

Goźdź-Roszkowski, Stanislaw, and Gianluca Pontrandolfo. 2017. *Phraseology in Legal and Institutional Settings: A Corpus-Based Interdisciplinary Perspective*. London: Routledge.

Halliday, Michael Alexander Kirkwood. 1994. Spoken and written modes of meaning. In *Media Texts: Authors and Readers*. Clevedon: Multilingual Matters Ltd., pp. 51–73.

Halliday, Michael Alexander Kirkwood, and Christian MIM Matthiessen. 2013. *Halliday's Introduction to Functional Grammar*. London: Routledge.
Jakubíček, Miloš, Adam Kilgarriff, Vojtěch Kovář, Pavel Rychlỳ, and Vít Suchomel. 2013. The tenten corpus family. Paper presented at 7th international Corpus Linguistics Conference, Lancaster, UK, July 23–26. pp. 125–27.
Kilgarriff, Adam. 2009. Simple maths for keywords. Paper presented at the Corpus Linguistics Conference 2009 (CL2009), Liverpool, UK, July 20–23. vol. 6.
Luttermann, Karin, and Jan Engberg. 2018. Vermittlung rechtlichen wissens an kinder und jugendliche im internet und in broschüren. In *Popularization and Knowledge Mediation in the Law/Popularisierung und Wissensvermittlung in Recht*. Münster: LIT Verlag, pp. 85–115.
Maley, Yon. 1994. *The Language of the Law*. London: Longman, vol. 1, pp. 11–50.
Masson, Michael E. J., and Mary Anne Waldron. 1994. Comprehension of legal contracts by non-experts: Effectiveness of plain language redrafting. *Applied Cognitive Psychology* 8: 67–85. [CrossRef]
Plain English Campaign. 2022. How to Write in Plain English. In *Plain English Campaign Drafting Resources*. New Mills: Plain English Campaign. Available online: www.plainenglish.co.uk/free-guides.html (accessed on 21 May 2022).
Pontrandolfo, Gianluca. 2021. National and eu judicial phraseology under the magnifying glass: A corpus-assisted analysis of complex prepositions in spanish. *Perspectives* 29: 260–77. [CrossRef]
Pontrandolfo, Gianluca. 2023. The importance of being patterned. In *Handbook of Terminology*. Amsterdam: John Benjamins Publishing Company, pp. 124–150.
Preite, Chiara. 2016. La vulgarisation des termes juridiques et la construction d'un savoir («que» faire) chez le grand public. *REPÈRES-DORIF* 10: 1–9.
Preite, Chiara. 2018. Stratégies dialogiques et transmission du savoir juridique dans le site du ministère de la justice française. In *Popularization and Knowledge Mediation in the Law/Popularisierung und Wissensvermittlung im Recht*. Münster: LIT Verlag, vol. 9, p. 149.
Rossetti, Alessandra, Patrick Cadwell, and Sharon O'Brien. 2020. "The Terms and Conditions Came Back to Bite": Plain Language and Online Financial Content for Older Adults. In *International Conference on Human-Computer Interaction*. Berlin and Heidelberg: Springer, pp. 699–711.
Tiersma, Peter M. 2000. *Legal Language*. Chicago: University of Chicago Press.
Turnbull, Judith. 2018. Communicating and recontextualizing legal advice online in English. In *Popularization and Knowledge Mediation in the Law/Popularisierung und Wissensvermittlung im Recht*. Münster: LIT Verlag, pp. 201–22.
Williams, Christopher. 2004. Legal English and plain language: An introduction. *ESP Across Cultures* 1: 111–24.
Williams, Christopher. 2015. Changing with the Times: The Evolution of Plain Language in the Legal Sphere. *Revista Alicantina de Estudios Ingleses* 28: 183. [CrossRef]
Williams, Christopher. 2022. *The Impact of Plain Language on Legal English in the United Kingdom*. Abingdon: Taylor & Francis.

languages

MDPI

Article

Who's Really Got the Right Moves? Analyzing Recommendations for Writing American Judicial Opinions

Mary C. Lavissière [1,*] and Warren Bonnard [2]

[1] Centre de recherche sur les identités, les nations et l'interculturalité, Nantes Université, CRINI, UR 1162, F-44000 Nantes, France
[2] ATILF, CNRS, Université de Lorraine, UMR 7118, F-54000 Nancy, France; warren.bonnard@univ-lorraine.fr
* Correspondence: marycatherine.lavissiere@univ-nantes.fr

Abstract: There is little linguistic research on the structure of judicial opinions from a discourse analysis perspective. There are, however, many professional resources about writing judicial opinions. This paper contributes to genre theory and linguistics of languages for specific purposes by proposing a role for professional writing advice. We also construct a typology of macrostructures proposed by professionals and compare them to the move structure of authentic judicial opinions. Our results show that, in terms of large discourse units, professional resources and move analysis seem to converge. Professional resources, however, do not describe the variation that may be observed in authentic documents. In this way, corpora of professional advice may contribute to a deeper understanding of how a discourse community represents its own genres.

Keywords: genre theory; move analysis; English for Specific Purposes; legal English; case law; annotation; corpus linguistics

Citation: Lavissière, Mary C., and Warren Bonnard. 2024. Who's Really Got the Right Moves? Analyzing Recommendations for Writing American Judicial Opinions. *Languages* 9: 119. https://doi.org/10.3390/languages9040119

Academic Editors: Julien Longhi and Nadia Makouar

Received: 16 November 2023
Revised: 11 March 2024
Accepted: 12 March 2024
Published: 26 March 2024

1. Introduction

"Does exploring the structure of opinions have any use?" (Leubsdorf 2002). In this article, we propose that it does. We agree that students of English for Legal Purposes (ELP) and legal professionals must have "at least an implicit understanding of this structure by learning to read an opinion as an opinion, rather than as some other kind of composition" (Leubsdorf 2002, p. 447). We also believe that linguistic analysis of these crucial documents is necessary for linguistics and for society. This paper is a first step toward these applied linguistic issues.

We are interested in judicial opinions in a common law system, and, in particular, in the decisions of the appellate and Supreme courts in the American judicial system. Regulators in the American federal and state systems do not advocate a universal format or structure. Kahn (2016, p. 5) reminds us that "it is not written anywhere that the court must issue an opinion; there are no rules requiring an opinion to take a certain form", making this type of communication "an unexpectedly complicated and subtle genre" (Leubsdorf 2002, p. 451). Its complexity leads to a need for guidance in the legal community of practice. Hafner (2014), for example, has pointed out that novice lawyers (students) aspire to demonstrate their writing expertise in a specific professional legal genre by adapting the codes and rhetorical choices made by experts in the discourse community. Vance (2011) argues that the legal community of practice, both professional and academic, has attempted to bring guidelines to opinion writing, with recommendations in textbooks for professionals and in the creation of legal writing courses for students preparing for a career as a clerk and then a judge.

Nevertheless, few sources of professional advice cited deal with aspects of structuring and organizing information in an opinion (Vance 2011). Instead, they deal with issues of professional ethics and the context in which clerks and judges operate. When the sources do address the writing process, they deal mostly with issues of personal style (with an

emphasis on "Plain English" recommendations), editing, or formatting. Hartig and Lu (2014, p. 88) summarize the issue as follows: "the majority of textbooks currently available on professional legal writing are not grounded in research-based descriptions of the genres that students are expected to produce".

The exceptions to the aforementioned trend of focusing on lower-level language matters are rare. Maley (1985), writing from a legal perspective, made a pioneering prototype for opinion structure based on his analysis of a single judgment. He claimed that the generic elements of opinions followed the FIRCO structure: "Facts, F, an account of events and/or the relevant history of the case; Issues, I, either of fact or of law; Reasoning, R; Conclusion, C, the principle or rule declared applicable for the instant case, and Order or Finding, O" (Maley 1985, p. 160). Bhatia (1993), writing from a discourse analysis perspective, depicts a similar overall structure, but emphasizes the interaction between legally significant facts and applicable law when reaching legal conclusions in judgments. Cheng and Sin (2007) and Gozdz-Roszkowski (2020) adopt Swalesian *move analysis* (Swales 1990). The former compares American judicial opinions to Chinese opinions and finds the following moves: heading, summary, facts and issues in dispute, arguments/discussion, decision/conclusion. The latter studies the structure of judges' justifications in the written decisions of the Polish Constitutional Tribunal, finding the following moves: object of constitutional review and constitutional issue, evaluating the admissibility of application based on pre-established criteria, reconstructing standards of review, evaluating the (non)compliance of a normative act with the Constitution, and evaluating the effect of ruling. From a natural language processing perspective, Kalamkar et al. (2022) identify and annotate for twelve *rhetorical roles* in a corpus of Indian judicial opinions: preamble, facts, ruling by lower court, issues, argument by petitioner, argument by respondent, analysis, statute, precedent relied, precedent not relied, ratio of the decision, ruling by present court, and one neutral category.

Given these different types of literature, we ask three research questions: (1) what structure for judicial decisions do professional manuals about writing opinions propose?; (2) how does the structure of judicial decisions in a corpus annotated using Swalesian theory compare to the prescriptive descriptions found in professional manuals?; and (3) how is the professional literature integrated into genre theory?

This paper is structured as follows: first, we review the literature about genre theory and move analysis, about the general role of expert advice in genre theory, and about legal writing manuals more specifically; second, we present our methodology for analyzing a set of legal writing manuals and for annotating a sample of judicial decisions with Swalesian discourse analysis. We then compare the results of our analyses of the manuals and the corpus. We discuss our results in light of the literature, and we conclude.

2. Materials and Methods

2.1. Theoretical Framework

2.1.1. Genre Theory

Before examining sources of professional advice about how to write a judicial opinion, we survey the literature on writing from the genre analysis perspective. Much of this literature was originally motivated by attempts to teach composition in humanities programs or writing skills in the framework of language acquisition or language for specific purposes. Genre is a useful concept for teaching writing because it covers expectations about the form and content of an instance of communication, especially written communication. While genre has long been an object of reflection in studies of rhetoric and of literature, genre theory as a framework for analyzing language used in professional context developed more recently in what have been known as the Australian school, the American school, and the school related to language for specific purposes (LSP) to which Swalesian analysis belongs.

Each school places a different weight on the social or communicative function and on structure in its definition of genre. New Rhetoric, the American school, which renewed interest in classical approaches to rhetoric while introducing concepts from social sciences and pragmatics, emphasizes the functional nature of genres over the formal aspects (Miller

1984). Miller (1984) also famously defined genres as a "social action" in the title of her article. Miller (1984, p. 159) states that genres are "typified rhetorical actions based in recurrent situations". Second, in systemic functional linguistics, the Australian school, there is also a focus on function, but it additionally highlights the importance of structure in defining genre as "staged goal-oriented social process" (Martin 2009). Third, in an LSP perspective, Swales (1990, p. 58) defines genre as "a class of communicative events, the members of which share some set of communicative purposes. [. . .] These purposes are recognized by the expert members of the parent discourse community and thereby constitute the rationale for the genre. This rationale shapes the schematic structure of the discourse". Swales' definition focuses more clearly on structure as related to community expectations and is the most widely used genre framework. Legal language is also more prone to respect a given structure because of the risk of litigation if documents are deemed not to conform to standards (Hiltunen 2012). For this reason, we adopt a Swalesian approach to studying the structure of judicial opinions.

Swalesian discourse analysis (Swales 1990, 2004; Moreno and Swales 2018) allows for studying genre at a meso level, between the whole document level and the lexico-grammatical level. Swales (1990) developed the framework in order to teach students of English for academic purposes (EAP) to learn the smaller units involved in creating a research article. The units defined by Swalesian analysis are called *moves* and *steps*. Moves are more abstract units, defined as "discoursal or rhetorical units performing coherent communicative functions in texts" (Swales 2004, pp. 228–29). Steps are concrete in that they are "text fragments" (Moreno and Swales 2018, p. 40) that "primarily function to achieve the purpose of the move" (Connor et al. 2007, p. 24) to which they belong.

Move analysis provides a framework for studying how language interacts with social expectations in a community of experts at multiple levels of genre analysis. Researchers carrying out move analysis have highlighted that the analysis cannot uniquely rely on a corpus of specialized documents, but implies interaction with the professional community itself (Tarone et al. 1998). For this reason, Swalesian move analysis has integrated feedback from these experts in three ways (Moreno and Swales 2018). First, before undertaking a move analysis in a given professional field, professionals may be surveyed about the types of documents that are crucial to their work. Second, experts may be asked to provide examples of typical documents for a given genre. Third, experts may interact with analysts to validate the final annotation schemes proposed (Moreno and Swales 2018, p. 41) "given their deeper knowledge of the text subject matter and their stronger intuitions regarding the typical rhetorical structure and language used in good papers in their fields". However, given that most of the literature using move analysis is published on research articles and, as such, the role of researcher and professional is not clearly distinguished, there is little research on how professional advice for producing written documents is treated in genre theory.

2.1.2. The Role of Expert Literature in Genre Theory

In short, we ask what role genre theory has reserved for expert advice in the form of manuals and articles rather than more spontaneous consultation of experts. As a pioneer in the analysis of legal discourse genres, Bhatia suggested that the study of legal genres should include a review of the related literature, and in particular professional literature (Bhatia 1993). However, increasing access to corpora of authentic legal documents has brought corpus-based studies on discourse to the forefront of genre studies. Our review of the scientific literature indicates that a comparison of professional literature to authentic corpora is lacking in English for Legal Purposes.

On one hand, existing literature focuses instead on what genre theory can bring to professionals in terms of instruction materials (Tribble 2009). This is, however, almost exclusively for students learning English for Academic Purposes (EAP). In this context, Tribble (2009) identifies three different traditions that EAP teachers use to teach writing to their learners. Particularly, Tribble (2009) identifies the "Social/Genre" tradition, which

features analyzing texts and discourse of a specific genre through its structural and lexico-grammatical elements (move analysis). On the other hand, as pointed out by Hyland (2012): "unlike much of the academic writing research, however, a great deal of professional writing research has been motivated less by pedagogical concerns than by the desire to gain an understanding of how people communicate effectively and strategically in organizations" (Hyland 2012, p. 104). In other words, other than EAP, research results coming from genre theory are not directly informing professionals about writing; professional writing advice, in turn, is not generally informing genre theory, with the exception of research writing. Concerning research writing, Norman (2003) questioned the instructions found in scientific-style manuals, in particular the instructions recommending a uniform terminology to designate constant entities in the same text. His study of a corpus of authentic documents concluded that these instructions were respected. Yang and Pan (2023), again based on a corpus of authentic documents, determined that the recommendations concerning the use of informal elements in writing generally corresponded to writers' practices.

The rift, with few exceptions, between professional writing advice and academic studies of specialized discourse is contradictory in the light of the importance of the concept of *discourse community* (Tribble 2015, p. 442) in genre theory. One of the scholars who adopts this term, Swales (1990, 2016), reminds us of the social nature of a discourse community, within which certain groups create or influence their own discursive practices as they develop conventions that meet their communication needs. Disciplinary experts are the most influential and attempt to direct communication to what they perceive as the needs of the discourse community. In return, novice members of the discourse community develop their writing expertise through their interaction with competent participants of the same community. Bhatia (2004, p. 165) highlights the high degree of interrelation between genre knowledge and what he defines as professional expertise, and asserts that the former "seems to be the key to pragmatic success in the use of language in wide-ranging professional contexts". In this way, it is through the concept of discourse community that expert advice about genre production is integrated into the definition of a particular genre and, by extension, into the notion of genre itself.

However, none of the literature, in the American context, has attempted to compare the extensive advice given about legal writing, and, in particular, the writing of judicial opinions, to actual legal documents. This is perhaps because of the singular place that writing and written documents have in law. While it is assumed that the foundations of professional writing, be it letters, emails, reports, or formal speeches, are learned during secondary and higher education, it cannot be assumed that students know how to write a brief or a judicial opinion without specific training. The singularity of these genres, their importance, and their complexity has led to a corpus of professional advice on legal writing that has yet to be fully incorporated into genre theory. We survey some of the characteristics of this literature in the following section.

2.1.3. Manuals

Manuals and professional articles, as argued in the previous section, can be viewed as a part of specialized communication, written by professional experts for professional novices with the objective of sharing their genre knowledge of the judicial opinion. Since their pragmatic aim is to 'tell how to do' in the course of a succession of obligatory or optional steps, they can be classified as procedural or instructional texts. Adam (2001) identifies several non-exclusive subsets under this umbrella name:

- Regulatory texts, which aim to regulate the behavior of one or more individuals;
- Programmer texts, where a speaker–programmer who is competent in a domain transfers their know-how to a reader–actor through the description of a process to be executed;
- Instructional–prescriptive texts that directly prompt action;
- The injunctive–instructional texts, similar to the previous ones in that they set up injunctive instructions;

- Advisory texts;
- Receipt texts.

According to Aouladomar and Saint-Dizier (2005), procedural texts use arguments based on the principles of rhetoric to convince the addressee to perform the prescribed actions. These texts, therefore, appeal to concepts of logic, but also to the emotions of their readers.

To sum up our literature review, we find three gaps in the literature about specialized genres, especially legal genres. First, the role of professional advice in the form of manuals is not well-defined in genre analysis. Second, there is no summary, neither in the legal nor in the linguistic literature about the overall structure proposed by the manuals and professional articles. Thirdly, a comparison between professional writing resources and an academic study of American judicial opinions is lacking. We propose a methodology for answering these questions in the following section.

2.2. Methodology

2.2.1. Manuals

For the analysis, we selected eleven different sources, based on their relevance and accessibility. Out of the 45 resources for judicial writing presented by Vance (2011), we collected those that were either available online or in French libraries. We then excluded the sources that did not provide recommendations or information on the actual or advisable structure of judicial opinions. In the end, seven resources presented by Vance (2011) were integrated into our analysis. These include two journal articles, two manuals for professionals, and three books addressing issues of legal writing and opinion writing. Besides Vance's (2011) sources, our own review of the existing literature about judicial writing led us to analyze three additional documents: three books intended to facilitate lawyers' understanding of judges' opinions. The sources we analyzed mostly concern appellate opinions in the US judicial system, although some of them are more general and include district court opinions, while one of them is more specific and addresses the question of Supreme Court opinions. Table 1, below, offers an overview of the sources and a more detailed description of manuals and articles reviewed can be found in Appendix A:

Table 1. Overview of the judicial manuals under study.

Reference	Descriptive /Prescriptive	Target Audience	Level of Detail	Author's Expertise
(Douglas 1983) "How To Write a Concise Opinion"	Prescriptive	Judges	Low	Judge
(Federal Judicial Center 2020) *Law Clerk Handbook*	Prescriptive	Law clerks	Low	Judges (judicial agency)
(Klein 1995) "Opinion Writing Assistance Involving Law Clerks: What I Tell Them"	Prescriptive	Law clerks	Low	Judge
(McKinney 2014) *Reading Like a Lawyer*	Descriptive	Students /lawyers	Low	Scholar
(van Geel 2009) *Understanding Supreme Court Opinions*	Descriptive	Students /lawyers	Low	Scholar
(Armstrong and Terrell 2009) *Thinking Like a Writer: A Lawyer's Guide to Effective Writing and Editing*	Prescriptive	Lawyers	Low	Scholars
(Federal Judicial Center 2013) *Judicial Writing Manual*	Prescriptive	Judges	High	Judges (judicial agency)

Table 1. *Cont.*

Reference	Descriptive /Prescriptive	Target Audience	Level of Detail	Author's Expertise
(Aldisert et al. 2009) "Opinion Writing and Opinion Readers"	Prescriptive	Judges/law clerks	High	Judge
(Aldisert 2012) *Opinion Writing*	Prescriptive	Judges/law clerks	High	Judge
(Sheppard 2008) "The 'Write' Way: A Judicial Clerk's Guide for Writing for the Court"	Descriptive /prescriptive	Students /lawyers	High	Scholar
(Leubsdorf 2002) "The Structure of Judicial Opinions"	Descriptive	Students /lawyers	High	Scholar

As mentioned in our introduction, our purpose is to compare the move structure of judicial opinions to the structure that is described and recommended in the resources created by the legal community. However, legal professionals do not describe legal opinions in terms of moves. The nature of the structure and linguistic observations put forth by the legal community about the communicative features of judicial opinions is heterogenic. Therefore, we will not refer to the categories established in our exploration of the professional literature as moves or steps. Instead, we use the term *macro-divisions*; these are defined using 1–5 below. We establish an inventory of potential macro-divisions as described by the resources about judicial opinion writing.

To establish this inventory, we used the following methodology:

1. In the manuals, we identified sections that indicate a description of the outline, the format, or the structure that opinions usually have or should have;
2. Within these sections, each specific feature described in at least one of the sources was then considered a potential macro-division of a judicial opinion;
3. After examining all the resources, we then counted the number of occurrences of a potential macro-division;
4. Additionally, we found lower level divisions inside the macro-divisions identified in phase 2 of our methodology. For example, some of the sources simply propose that opinions include a section in which the court justifies the final decision without giving details about its content. Others provide details about the stages involved in the justification (see Figure A1 for examples);
5. Before presenting the results of this exploration, we classified our sources according to different analytic criteria: descriptive/prescriptive, author's expertise, target audience, and level of detail.

Our overview of the sources led us to classify the macro-division into different groups. First, we carried out an analysis of the macro-divisions that we found in sources with a low level of detail and then to those found in sources with a high level of detail. Sources with a low level of detail offer only very brief descriptions of the communicative features that should make up a judicial opinion, usually in one page or even less for the whole account of the structure of a judicial opinion. Sources with a high level detail offer longer descriptions of the divisions, usually, more than three pages. When a macro-division was present in all or virtually all the sources of a group, we also provided the position in which their author said they appear or should appear in a judicial opinion. One exception was the source "The Structure of Judicial Opinions", which does not explicitly state the order of appearance of the opinion's main elements.

The authors' expertise criterion seemed to correlate with the descriptive/prescriptive criterion. For this reason, we then compared the macro-divisions of the source group *Judges* vs. the group *Scholar*. This distinction reflects the possible different views towards the structure of the judicial opinion within the legal community.

2.2.2. Move Analysis

The move analysis was carried out on a sample corpus of Supreme Court of the United States (SCOTUS) majority opinions. We describe the two phases of annotations and the construction of a representative corpus in the following paragraphs.

The annotation took place in two phases. In the first phase, we based our methodology on Moreno and Swales's (2018) move analysis of the *discussion* section of research articles. In this, steps are annotated based on functional rather than formal criteria: "A step is a text fragment containing 'new propositional meaning' from which a specific communicative function can be inferred 'at a low level of generalization by a competent reader of the genre'" (Moreno and Swales 2018, p. 49). As per its linguistic instantiation, "[a] step can be realised by a proposition, a proposition complex or an even larger fragment of text". (Moreno and Swales 2018, p. 48). However, because the annotations are part of a larger project (Lexhnology ANR-22-CE38-0004) that includes machine learning, we first had to impose a clear and coherent segmentation based on formal units. We chose the unit of a grammatical sentence. We assigned a communicative function, a step, to each unit.

For the annotation itself, we used the free annotation software Taguette (Rampin and Rampin 2018) to carry out an exploratory study of 5 historical landmark opinions in SCOTUS case law (see Appendix B). We annotated a shared version of each opinion. This phase resulted in more than 100 steps for the 5 opinions. Following Moreno and Swales (2018), we also proposed 10 prototypical moves (see Figure A3 in Appendix B) based on observations of patterns in some of the opinions, on the manuals, and on how we thought the moves should logically be constituted. We especially focused on transitions that signaled the openings and closings.

We then decided to test the moves and steps from the first phase on a more representative corpus of SCOTUS majority opinions. The corpus was constructed by consortium partners in Lexhnology (see Acknowledgements for details). We used the SCOTUS Opinions (Fiddler 2020) corpus available online. SCOTUS Opinions was itself taken from the website Court Listener (CourtListener 2024) and enriched with metadata about the cases. As majority opinions are long and the annotation process is complicated, we wanted to create a sample corpus with the smallest number of opinions that would optimize representativity in terms of date, length, and theme. We, therefore, created a representative corpus for the majority opinions from 1945 to 2020 using two criteria: the justice listed as author of the majority opinion and theme. Author was chosen because this variable covers the variation linked to date and length of the opinions. As for the themes, they were constructed with a K-means clustering algorithm (K = 18) using term frequency * inverse document frequency. This process allowed us to group our data into thematic homogeneous groups, resulting in 18 different themes. Finally, we chose a threshold of the first 4 most productive justices for each theme, as this gave us the best representativity for the smallest number of opinions. The total number of opinions in the sample corpus was 18.

This representative corpus was used in the second phase of annotation. We tagged the full sample of 18 opinions described in the preceding paragraph. Each opinion was annotated separately by one of the authors of this article without consultation. After each opinion was annotated, we then discussed each segment to resolve differences in annotation. At the end of this process, we reduced the initial list of more than 100 steps to 34 steps (see Figure A4 in Appendix B). These were validated by an external legal expert. All 18 opinions were reannotated with the final set of annotations. The current Cohen's Kappa for the annotation scheme and process is 0.66, which indicates that we have achieved a good level of coherence when we annotate the same opinion separately. In terms of distribution, Figure 1 below shows the percentage of annotation agreement between the two annotators on the left. On the right, Figure 1 shows how frequently an annotation appears. To give a few examples, some annotations, such as *granting certiorari* are infrequent because they appear only once in each opinion; however, they contain fixed language and are easy to identify. Other annotations, such as *recalling a SCOTUS opinion,* appear frequently, but

are more difficult to identify because they are more closely related to other interpretation annotations, such as *stating the Court's reasoning*.

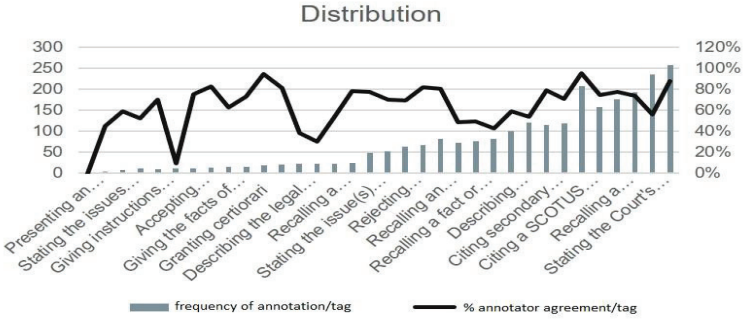


Figure 1. Percentage of annotator agreement according to frequency of individual annotation.

The frequency of appearance per annotation was obtained by counting the number of occurrences of each annotation for each annotator. We then harmonized these results by averaging the appearance frequencies per annotation for each annotator. The annotator agreement per annotation was obtained by counting the number of times both annotators had chosen to annotate the same segment with the annotation in question. This number was then divided by the number of occurrences of the annotation by each annotator. We then harmonized these results by averaging the annotator agreements per annotation and per annotator.

Importantly, the prototypical moves identified in the first annotation phase did not appear in the representative sample. This points to wide variation of SCOTUS opinions in terms of larger discourse units. We plan to investigate the move level of macro-divisions in a later study using machine learning to identify regular patterns of steps. For this paper, however, our corpus-based study focused on the steps that were manually annotated. This choice is in line with the methodology proposed by Moreno and Swales (2018, p. 44), who conclude after reviewing the move analysis literature that "steps may be better indicators of shared psychological realities than moves and that annotating [. . .] sections for their steps before conceptualising the moves might help us to arrive at a clearer picture of what is happening [. . .]". In our methodology, the steps were then compared to the macro-divisions found in the expert literature.

3. Results

3.1. Manuals

A summary of our analyses of the manuals is presented in Figures A1 and A2 (see Appendix A), which indicate each source in the left-hand column, the presence or absence of the rhetorical divisions identified from that same source or from others included in our scope of research. Figure A1 separates the sources that evoke the structure of judicial decisions in detail from those that do not enter into detail. Figure A2 separates sources written by judges from those written by academics.

The figures show that five macro-divisions are found in almost all the sources analyzed:

- Introduction/Orientation Paragraph(s);
- Facts;
- Issues to be decided/legal questions;
- Justification of the decision of the court;
- Final action taken by the reviewing court.

In particular, for sources with a low level of detail, these divisions are sometimes the only structure cited by the experts. For example, Justice Charles Douglas (1983, pp. 4–6) describes them as the "five constituent parts of an opinion", without developing their content further other than in a few sentences about the *introduction*, which he describes as

follows: "the nature of the action and how it got to the appellate court". These divisions echo the FIRCO structure proposed by Maley (1985). However, our observations highlight the almost systematic presence of a section introducing the opinions. In addition, we note that while the concluding part of the opinions ('Final action taken by the reviewing court') seems to omit the letter O of Maley's model, in reality, when we look at sources that are more generous in their descriptions or recommendations, we notice that the announcement of the judgment is often followed by instructions to the lower courts, particularly when the case is remanded. This shows a fit with the FIRCO model.

Thus, according to legal professionals, the five main divisions seem to represent the backbone of any judicial decision. This also echoes the Greco-Roman model of the art of persuasive rhetoric presented by Aldisert (2012):

- Exordium, i.e., the introduction that presents the major questions of the case (How? What? Who? When? Where?);
- Divisio, or the announcement of the division of the case according to the legal issues to be discussed;
- Narratio, the recounting of the facts of the case;
- Confirmatio, i.e., the presentation of evidence to analyze the parties' arguments on the points of law;
- Peroratio, the final legal conclusion of the case.

According to Aldisert et al., "these five parts form the structure of every well written opinion. Each is absolutely essential" (Aldisert et al. 2009, p. 24). As these parts are found in almost all the sources analyzed, we were also able to compare their order of appearance in the texts according to their authors. This analysis is of interest only for the intermediate parts, but it shows that there is no consensus on the question of presenting facts before legal issues. Indeed, excluding Leubsdorf's (2002) article, for which the order of divisions is omitted, facts are presented before legal issues in half the sources observed. This quotation from Aldisert et al. (2009) may shed some light on divergent points of view: "The statement of issues usually should precede the narration of the facts. [...] This does not mean that the statement of issues must always precede the statement of facts in the final draft of the opinion. That may depend on style" (Aldisert et al. 2009, p. 28).

This is, therefore, a matter of personal preference, and may also depend on the type of case. In some highly procedural cases, the facts in dispute are the very problem to be solved. Thus, some authors do not attribute a formal rule to this question, and speak of conventions that evolve over time. For McKinney (2014, p. 23), the presentation of legal issues must take place "someplace in the opinion". Within the lower level divisions, only the *introduction/orientation paragraphs* and *justification of the decision of the Court* sections are the subject of significant clarification regarding their content. The legal professionals emphasize the role of the introductory paragraphs in setting the scene, focusing above all on the presentation of the parties and the procedural elements that led to the trial before the adjudicating court. Most authors also emphasize the introductory section's function of anticipating information, which they believe should announce the legal issue and the final judgment before they are repeated later in the opinion.

The *justification* section, on the other hand, contains few elements, according to the sources studied, which indicate that the section must apply the law to the facts. It should be noted, however, that the high-level sources systematically include an assessment and then a rejection of the losing party's claims and arguments in the present case.

Finally, a comparison of the sources written by judges and those written by scholars reveals that they differ above all in the introduction. Judges attach greater importance to this part, and, thus, recommend that the parties to the case be precisely identified. The anticipatory function of the introduction is also widely emphasized by the judges, who clearly distinguish themselves from the scholars by recommending that both the legal problem and the final judgment be announced in the very first lines of the opinion.

3.2. Move Analysis

As compared to the analysis of manuals, the move analysis shows that SCOTUS opinions vary widely in their structure. We grouped the 34 finalized step annotations (see Appendix B, Figure A4) into the categories shown in Figure 2 (below). It is important to emphasize that the categories are not moves, but a classification of types of steps: indeed, we believe that moves are situated at an intermediate level between the step-level and the category-level. Other differences also exist. The categories *legal question(s)* and *legal sources* are thematic rather than communicative. They are, thus, likely to appear throughout the text, interspersed amongst steps pertaining to another category. Furthermore, a third category, which is metadiscursive (*announcing function*), contains only one step and also appears throughout the text. The three remaining categories represent broad communicative functions. They are analogous to sections of research articles, i.e., *introduction*, *literature*, *methods*, etc. These are broad and very long sections that are generally considered to be at a higher level of discourse than moves in Swalesian analysis.

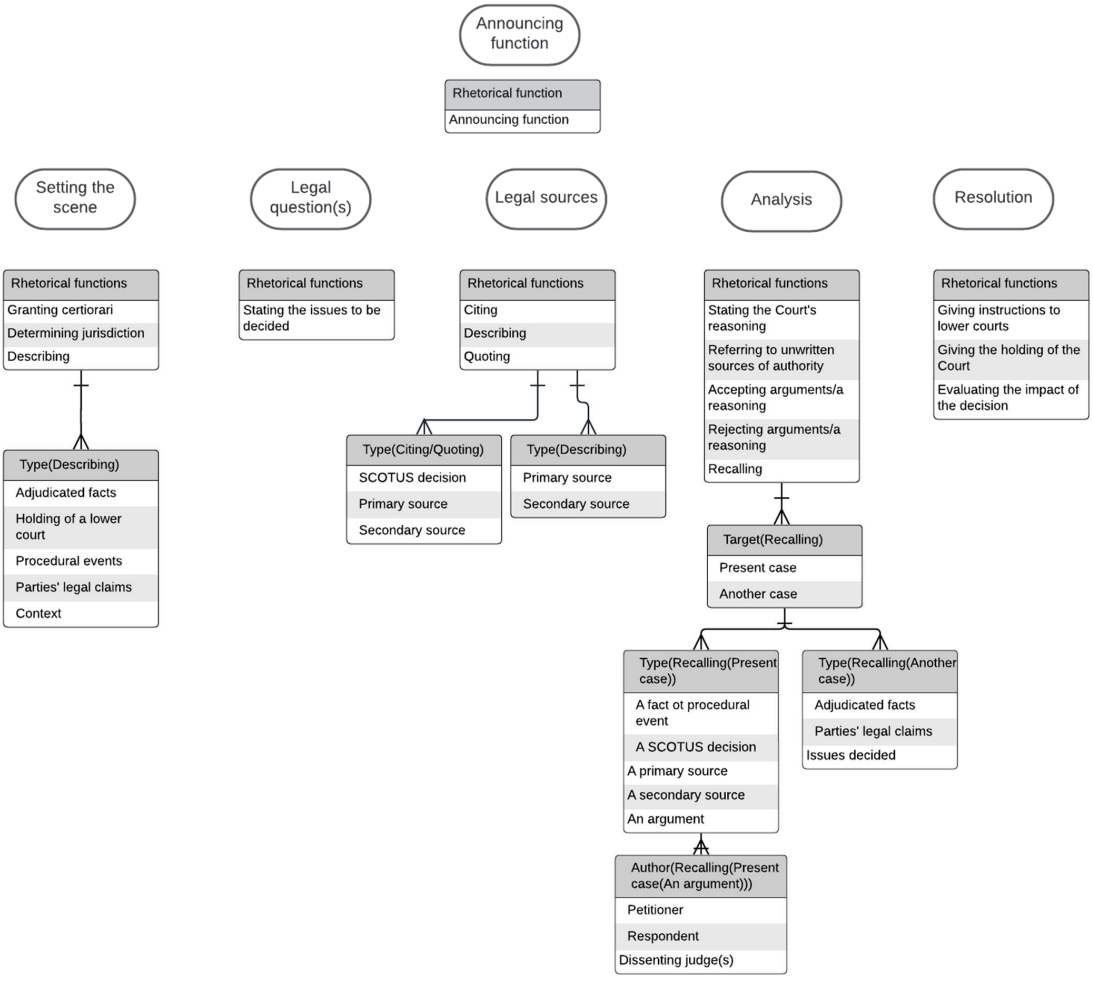


Figure 2. Categorization of steps.

The typology includes five larger categories:

1. Setting the scene: the Court's decisions usually consist of introductory paragraphs that serve to introduce the case to the reader. They, therefore, contain elements relating to the nature of the parties, their claims in the case, the material facts concerning them, and the course of the proceedings that led the Court to rule on the case as final jurisdiction;
2. Legal questions: the indication of the legal issue to be resolved by the Court generally comes at the end of the setting of the case. However, we have placed it in a separate category, since we have found that it can also be stated several times in the opinion, sometimes in ways that broaden or narrow the issue;
3. Legal sources: we have also placed sources of law in a separate category. Indeed, many sources of law, whether case law, legislation, or the Constitution, appear throughout the opinions. Sometimes, the judges clearly highlight a section in their opinion that describes the sources of law that will be discussed in the *analysis* section. However, whenever sources of law are evoked in support of the judges' arguments, we consider that they should be associated with the rhetorical functions of the *analysis* section;
4. Analysis: this category corresponds to the heart of SCOTUS opinions. It is generally the longest, beginning after the expository part and ending before the statement of the final decision. The text is argumentative in nature, and includes the Court's justification in response to the parties' arguments;
5. Resolution: this section reports on the resolution of the legal problem of the case by stating the final decision, following the argument of the majority. If the final judgment is mandatory, it may sometimes be accompanied by instructions for the lower courts and/or considerations of the impact of the decision on civil society.

In terms of the overall order in which the steps appear, some appear in a relatively fixed position in the text. In general, steps that appear in the category *setting the scene* appear first, followed by those in the *analysis*, and finally, *resolution*. Other steps, however, appear throughout the opinion. These include the metadiscursive step *announcing* function, the *issue*, and the steps related to dealing with other sources. Crucially, the largest number of steps relate to the *analysis* of the justices. They also represent the largest part of the opinions.

The annotation process and final annotation scheme make it clear that the macro-divisions proposed by the manuals and professional articles are not on the same level as steps. Overall, the macro-divisions proposed by professionals are based on long-standing rhetorical divisions that are larger than steps. Nor are they moves, as moves had not yet emerged from the representative sample corpus. This is because there is wide variation in how the steps appear in opinions in the sample corpus. We also did not find clear transitions that would open and close moves in the sample corpus. We discuss our findings in detail in the following section.

4. Discussion

Our study contributes to the conversation about the role of expert advice in language for specific purposes. We structure this discussion around our three research questions, starting from the most concrete and moving to the more theoretical.

4.1. The Structure of Judicial Decisions According to Professional Manuals

The macro-divisions in the manuals include introduction, facts, issues, reasoning, and final decision. These divisions are similar to the divisions observed by some researchers, such as Bhatia (1993). According to Aldisert (2009), this organization is highly influenced by classical rhetoric studies.

4.2. Comparison of Professional Manuals and Authentic Documents

We also investigate how the content of professional manuals compares with authentic documents. These questions are relatively new to the disciplinary framework of legal English, despite Bhatia's (1993) earlier recommendations to carry out such comparisons. While this study is the first to adopt Bhatia's (1993) suggestions for legal English, such a

comparison exists in other specialized domains at a microlevel, as shown by the studies of Norman (2003) and Yang and Pan (2023) who found that expert advice was coherent with what was found in authentic scientific writing. As compared to the studies mentioned, we study a higher level of discourse and legal language. In our study of judicial opinions, the divisions we observe in prescriptions or descriptions in professional writing advice sometimes also correspond to those in the corpus of judicial opinions. In our corpus, we observe categories of steps that are similar to the five categories identified in the majority of manuals: *setting the scene* in our corpus study is similar to *introduction* in the manuals; *analysis* is similar to *reasoning*; and *resolution* is similar to *final decision*. These sections generally appear in the same order, in the professional manuals, academic research (Bhatia 1993; Kalamkar et al. 2022), and in our corpus study. Furthermore, *issue* is identified in all of these studies. Our study shows, however, that this division does not always appear in the same place.

One major difference in our corpus study is that we carry out an analysis at a lower level of discourse. At step level, for example, we find a variety of text fragments related to dealing with different "sources of discourse". This level of precision is not found in most manuals. The manuals, therefore, do not include the wide range of units and discourse strategies used to build the legal opinions, for example the variety of steps in the *analysis* category identified during our move analysis. In addition, the manuals present judicial opinions as having a stable structure. In a similar vein, the recommendations set out in the manuals are deliberately vague.

Legal professionals' effort to link the Aristotelian rhetorical model to the structure of opinions may be aimed at demonstrating continuity with accepted models of persuasion. According to this reasoning, accepted models may be legitimately reproduced and, as a consequence, are generally observed in legal opinions. In addition, by arguing that opinion writing is based on ancient and immutable rhetorical principles, legal professionals may avoid giving the impression that legal decisions are arbitrary.

The findings of this study, however, contradict the representation of SCOTUS opinions as uniform and stable. Instead, they suggest that the structure of SCOTUS opinions is highly variable. For example, no moves can currently be constructed from the steps because full annotation will be necessary to recognize, possibly with the help of machine learning, patterns of steps. This variation may be explained by the fact that there is no ready-made answer to justify a legal decision in the United States. The legal cases handled by American courts are highly diverse, and this diversity of legal facts necessarily calls for a tailor-made response from judges. It can also be seen as a desire to leave the field of interpretation open. According to Black et al. (2016), SCOTUS justices alter their opinion writing to improve compliance and to increase the general public's acceptance of their decisions. The same idea of strategic writing can be applied to communicative structure.

In sum, what characterizes SCOTUS decisions is the relative freedom to deploy arguments to achieve varying communicative purposes. A highly elaborate and standard framework for opinion-writing would reduce this freedom, especially because all American justices and judges do not necessarily adhere to the same school of legal theory or target the same audience members. SCOTUS opinions, for example, may be different from the opinions of lower appellate courts because SCOTUS opinions are covered by the press. SCOTUS justices also tend to have a distinct style, which may introduce even more variation into the structure of the opinions.

4.3. Contributions to Move Analysis and Genre Theory

Our study contributes nuances to move analysis and genre theory. We find that the divisions we observe in the opinions are more cyclical than moves in scientific articles. Discourse in case law is based on constructing logical arguments using different sources of discourse. Scientific articles, on the contrary, construct a more linear argument that clearly 'moves' in one direction. The popularity of the model developed by Swales lies in its ability to account for discursive configurations inherent to a given genre. These configurations are

not immutable, however, and when they are applicable, which is not always the case (see Maswana et al. (2015) in the hard sciences, and Lu et al. (2021) for the social sciences), they can be modified according to the needs of the writer.

Our corpus-based study of macro-divisions in SCOTUS opinions reveals that linear progression does not always match how judges write opinions. At the step level, we found cyclical movements based on interdiscursivity and intertextuality. Indeed, to support their arguments, judges often call on external sources of law, on their own discourse set out earlier in the judgment, or on the arguments of the losing party in order to reject them. This intermingling of external sources of discourse tends to disrupt the communicative unity of the judges' argument and the linear trajectory of the discourse's main thread. In some cases, such as Baldwin v. Reese (2004), the cyclical dimension of the discourse is explicitly present, as seen in Figure 3.

Figure 3. Example of cyclical distribution of *Analysis* steps in Baldwin v. Reese (2004).

Finally, we contribute to genre analysis by questioning the role of expert advice in studying genre. Our literature review and our studies show that (1) expert advice does not completely deviate from the structure of authentic corpora; (2) when there are a number of sources of expert advice, such as about opinion writing, they tend to converge; (3) expert advice tends to present a simplified representation of the documents described. For example, the variation we observe in the structure of SCOTUS opinions is not addressed in the manuals; (4) this may be motivated by the desire to give non-members of the discourse community the impression that the production of judicial opinions is unified and, thus, controlled, whereas, in practice, the lack of constraints about opinion writing give justices large margins of freedom to adopt variable patterns of justification; (5) expert advice in opinion writing tends to integrate traditions from larger persuasive language, such as Greco-Roman models; (6) expert advice on language should be a more prevalent object of study for linguistics and language for specific purposes. While this discourse does not represent the complexity of authentic corpora, collections of expert advice contribute to a discourse community's auto representation, which contributes nuance to genre analysis. They are also sources for studying interaction between a given specialized language and less specialized language, as the references to Greco-Roman rhetoric show in legal argumentation.

5. Conclusions

This study contributes to a richer vision of specialized genres and language, in which expert advice is both an additional source of information about specialized discourse and an object of study itself. Much research remains to be done, both on legal expert advice and on the opinions itself. One future perspective for genre research that our findings highlight is the representation of genres by their own discourse communities. In the case of

SCOTUS majority opinions, this aspect may be further researched once the annotation of the full corpus has been achieved. The full annotation will allow for a closer observation of trends of step organization and move construction. Interviews with American judges and other legal professionals may allow for comparing current perceptions of the discourse community about the structure of majority opinions with their actual structure. Future studies may include investigating courses about writing or reading judicial opinions. For the time being, however, the present study points to the complimentary information that may be found by analyzing both the professional literature about majority opinions and our corpus of SCOTUS opinions.

Author Contributions: Conceptualization, M.C.L.; methodology, M.C.L. and W.B.; validation, M.C.L.; formal analysis, M.C.L. and W.B.; investigation, M.C.L. and W.B.; resources, M.C.L. and W.B.; data curation, M.C.L. and W.B.; writing—original draft preparation, M.C.L. and W.B.; writing—review and editing, M.C.L. and W.B.; visualization, M.C.L. and W.B.; supervision, M.C.L.; project administration, M.C.L.; funding acquisition, M.C.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Agence Nationale De Recherche, grant number ANR-22-CE38-0004.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available because the research is ongoing.

Acknowledgments: We acknowledge the work of Nicolas Hernandez and Anas Belfathi (Nantes Université, LS2N, UMR 6004, F-44000 Nantes, France) in preparing the representative sample corpus of the SCOTUS Opinions corpus.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

- Aldisert (2012) and Aldisert et al. (2009): two of the sources are written by judge Ruggero J. Aldisert. At the time of publication, he was an appellate judge in the U.S. Court of Appeals for the Third Circuit. The first source is a monograph entitled *Opinion Writing* (Aldisert 2012). It provides guidance for professionals who aim to learn or to perfect their opinion writing. The second source by Aldisert is an article entitled "Opinion Writing and Opinion Readers" (Aldisert et al. 2009). It draws on the topics presented in the second edition of *Opinion Writing* "while specifically highlighting the relationship between opinion writing and opinion readers" (Aldisert et al. 2009, p. 4). These two sources offer the most detailed description of the structure of an opinion in terms of macro-divisions. *Opinion Writing* has an entire section dedicated to the "Anatomy of an Opinion", while "Opinion Writing and Opinion Readers" intends to "dissect the ideal structure of an opinion" (Aldisert et al. 2009, p. 4). These sources, thus, adopt a prescriptive view towards the professional community of judges and law clerks. They are based on the author's 30-year experience as an appellate judge.
- Douglas (1983): "How to write a concise opinion" (Douglas 1983) is a short article written by appellate judge Charles G. Douglas. It was published in the *Judges Journal*, suggesting a readership composed of a majority of judges. It adopts a very prescriptive view on how judges should write their opinion in order to reduce its length.
- Federal Judicial Center (2013, 2020): the handbooks *Law Clerk Handbook* and *Judicial Writing Manual* are comparable in that they are intended as manuals for professionals. Both are written by the same judicial agency, the Federal Judicial Center. They provide an overview of legal professionals' specific duties with a chapter focusing on research and legal writing. These two sources are prescriptive and provide the same type

of guidance. *Judicial Writing Manual*, however, offers a more detailed view of the structure that an opinion should have.

- Klein (1995): "Opinion Writing Assistance Involving Law Clerks" is an article written by judge Richard B. Klein. It specifically addresses law clerks and recommends guidelines to improve their opinion writing. These recommendations are drawn from Klein's own experience as a judge, as they "fit his personal style" (Klein 1995, p. 7). The article is relatively detailed and comments on lower levels of language as well.
- Sheppard (2008): "The 'Write' Way: A Judicial Clerk's Guide to Writing for the Court" is an article written by Jennifer Sheppard and published in the *University of Baltimore Law Review.* Jennifer Sheppard is an Assistant Law Professor at Mercer University School of Law. She addresses law clerks and law students and offers a fairly detailed approach to writing an opinion. Her article is both descriptive and prescriptive as regards to format. She presents excerpts from actual opinions and guidelines to clerks, based on what an opinion generally includes. Her description is not, however, informed by an identified corpus of opinions. Importantly, she states that "the format of an opinion may vary depending on the court or the case itself" (Sheppard 2008, p. 79).
- Armstrong and Terrell (2009*): Thinking like a Writer: a Lawyer's guide to effective writing and editing* has only a short section about judicial opinion writing. The rest of the book presents a large set of principles and tips that lawyers should use to improve their legal writing in general. These same principles are used to describe "the structure of a simple opinion" (Armstrong and Terrell 2009, p. 259). The description of macro-divisions is based on what the authors, drawing on their experience, consider logical and coherent for a judicial opinion.
- van Geel (2009): *Understanding Supreme Court Opinions* is a book intended for law students. Its author, T.R. Van Geel (professor of law and political science), says that it should supplement their constitutional casebook material. The audience can also include lawyers who wish to improve their understanding of SCOTUS opinions. The book adopts a descriptive approach. The organizational structure is referred to in terms of the most typical elements shared among Court's opinions.
- McKinney (2014): The book *Reading Like a Lawyer* was written by the law professor Ruth Ann McKinney. As the title suggests, its target audience is lawyers who want to improve their strategies for reading case law. The book includes a brief description of the structure of a judicial opinion, based on "conventions that have evolved over time" (McKinney 2014, p. 23). No further detail about the macro-divisions is given.
- Leubsdorf (2002): "The Structure of Judicial Opinions" takes a linguistic perspective when describing judicial opinions. This distinguishes the long article from the other sources studied in this article. Its author, law professor John Leubsdorf, describes these documents being examples of a complex genre that intertwines different voices and stories. Although it includes the main elements of an opinion in detail, the piece does not explicitly state how the information is organized within these elements, nor does Leubsdorf present the order in which law students or lawyers are expected to encounter the information within the main elements. Like most of the professional sources, Leubsdorf's analysis relies on his own practice and experience as a law practitioner.

Additional Information about the Expert Sources under Study

	Introduction/Orientation paragraph	Nature of the action and of the parties	Lower's court or agency's reasons for its decision	Parties' legal claims	Judgment/ holding from the present court	Relevant prior procedural events	Issue to be discussed	Basis of the court/lower court for jurisdiction	Facts	Interpreting the facts of the case	Issues to be decided/legal questions	Standard of review for each issue (or controlling legal principle)	Justification of the decision of the court	Application of the law to the facts	Tests	Rejection of losing side's contentions	Criticism of dissenters' arguments	Reliance on background facts	Final action taken by the reviewing court	Instructions to the lower court(s) (if case is remanded)
How To Write a Concise Opinion, (1983)	1	X	X			X	X		3		2		4	X					5	
Law Clerk Handbook, (2020)	1	X	X		X		X		2			3	4	X					5	X
Opinion Writing Assistance Involving Law Clerks: What I Tell Them, (1995)	1	X		X	X	X			3		2		4	X		X			5	
Reading Like a Lawyer, (2014)					4	1			2		3		5						6	
Understanding Supreme Court Opinions, (2009)				3		2			1	X	4		5	X	X	X	X	X	6	
Thinking Like a Writer: A Lawyer's Guide to Effective Writing and Editing, (2009)	1	X		X	X	X	X		2			X	3	X		X			4	
Judicial Writing Manual, (2013)	1	X			X	X	X	X	3		2	X	4			X		X	5	X
Opinion Writing- and Opinion Readers, (2009)	1	X	X		X	X	X	2	5		3	4	6	X		X			7	
Opinion Writing, (2012)	1	X	X		X	X	X	2	5		3	4	6	X	X	X			7	X
The "Write" Way: A Judicial Clerk's Guide for Writing for the Court, (2008)	1	X	X		X	X		X	2	X	3	4	5	X		X			6	X
The Structure of Judicial Opinions, (2002)	X		X	X		X			X		X		X	X		X	X	X	X	X

Figure A1. Candidate macro-divisions (in green) and micro-divisions: high level of detail (in grey) vs. low level of detail (in yellow) in the sources under study.

	Introduction/Orientation paragraph	Nature of the action and of the parties	Lower's court or agency's reasons for its decision	Parties' legal claims	Judgment/ holding from the present court	Relevant prior procedural events	Issue to be discussed	Basis of the court/lower court for jurisdiction	Facts	Interpreting the facts of the case	Issues to be decided/legal questions	Standard of review for each issue (or controlling legal principle)	Justification of the decision of the court	Application of the law to the facts	Tests	Rejection of losing side's contentions	Criticism of dissenters' arguments	Reliance on background facts	Final action taken by the reviewing court	Instructions to the lower court(s) (if case is remanded)
How To Write a Concise Opinion, (1983)	1	X	X			X	X		3		2		4	X					5	
Law Clerk Handbook, (2020)	1	X	X		X		X		2			3	4	X					5	X
Opinion Writing Assistance Involving Law Clerks: What I Tell Them, (1995)	1	X		X	X	X			3		2		4	X		X			5	
Judicial Writing Manual, (2013)	1	X			X	X	X	X	3		2	X	4			X		X	5	X
Opinion Writing- and Opinion Readers, (2009)	1	X	X		X	X	X	2	5		3	4	6	X		X			7	
Opinion Writing, (2012)	1	X	X		X	X	X	2	5		3	4	6	X	X	X			7	X
Thinking Like a Writer: A Lawyer's Guide to Effective Writing and Editing, (2009)	1	X		X	X	X	X		2			X	3	X		X			4	
Reading Like a Lawyer, (2014)					4	1			2		3		5						6	
Understanding Supreme Court Opinions, (2009)				3		2			1	X	4		5	X	X	X	X	X	6	
The "Write" Way: A Judicial Clerk's Guide for Writing for the Court, (2008)	1	X	X		X	X		X	2	X	3	4	5	X		X			6	X
The Structure of Judicial Opinions, (2002)	X		X	X		X			X		X		X	X		X	X	X	X	X

Figure A2. Candidate macro-divisions (in green) and micro-divisions depending on author's expertise (judges in pink, scholars in blue).

Appendix B

Ascertaining the Supreme Court's jurisdiction for the case
Introducing the case
Describing the facts and procedural history of the case
Analyzing arguments/reasoning from the lower courts
Analyzing an argument from the petitioner
Analyzing an issue for the present court
Criticizing arguments from the dissenters
Referring to other cases to justify the decision of the present court
Presenting the arguments of the Court
Stating the disposition of the case

Figure A3. Ten prototypical moves from first phase of annotation.

Accepting arguments/reasoning
Announcing function
Citing a SCOTUS decision
Citing primary sources (other than SCOTUS decisions)
Citing secondary sources
Describing primary sources (other than SCOTUS decisions)
Describing procedural events
Describing secondary sources
Describing the adjudicated facts
Describing the legal arguments made in another case
Describing the legal arguments made in the present case
Determining jurisdiction
Evaluating the impact of the decision
Giving instructions to the courts below
Giving the context
Giving the facts of another case
Giving the holding/ruling of a lower court
Giving the holding/ruling of the Court
Granting certiorari
Presenting an argument from the dissenters
Quoting a SCOTUS decision
Quoting primary sources
Quoting secondary sources
Recalling a SCOTUS decision
Recalling a fact or procedural event
Recalling a primary source (other than a SCOTUS decision)
Recalling a secondary source
Recalling an argument from the petitioner
Recalling an argument from the respondent
Referring to unwritten sources of authority
Rejecting arguments/a reasoning
Stating the Court's reasoning
Stating the issue(s) to be decided
Stating the issues decided in another case

Figure A4. Final set of step annotations after the second phase of annotation.

References

Adam, Jean-Michel. 2001. Types de textes ou genres de discours? Comment classer les textes qui disent de et comment faire? *Langages* 35: 10–27. [CrossRef]
Aldisert, Ruggero J. 2009. *Opinion Writing*, 2nd ed. Bloomington: AuthorHouse.
Aldisert, Ruggero J. 2012. *Opinion Writing*, 3rd ed. Durham: Carolina Academic Press.

Aldisert, Ruggero J., Meehan Rasch, and Matthew P. Bartlett. 2009. Opinion Writing and Opinion Readers. *Cardozo Law Review* 31: 43.
Aouladomar, Farida, and Patrick Saint-Dizier. 2005. Towards Generating Procedural Texts: An Exploration of Their Rhetorical and Argumentative Structure. Paper present at the Tenth European Workshop on Natural Language Generation (ENLG-05), Aberdeen, UK, August 8–10.
Armstrong, Stephen V., and Timothy P. Terrell. 2009. *Thinking Like a Writer: A Lawyer's Guide to Writing and Editing*, 3rd ed. New York: Practsising Law Institute.
Baldwin v. Reese. 2004. 541 US 27. Washington, DC: U.S. Supreme Court Center.
Bhatia, Vijay K. 1993. *Analysing Genre: Language Use in Professional Settings*. London: Longman.
Bhatia, Vijay K. 2004. *Worlds of Written Discourse: A Genre-Based View*. London: Bloomsbury Academic.
Black, Ryan C., Ryan J. Owens, Justin Wedeking, and Patrick C. Wohlfarth. 2016. *U.S. Supreme Court Opinions and Their Audiences*. Cambridge: Cambridge University Press.
Cheng, Le, and King-kui Sin. 2007. Contrastive Analysis of Chinese and American Court Judgments. In *Language and the Law: International Outlooks*. Edited by Krzysztof Kredens and Stanislaw Goźdź-Roszkowski. Berlin: Peter Lang, pp. 325–56. Available online: https://www.peterlang.com/view/title/51172 (accessed on 10 November 2023).
Connor, Ulla, Thomas A. Upton, and Budsaba Kanoksilapatham. 2007. Introduction to move analysis. In *Discourse on the Move: Using Corpus Analysis to Describe Discourse Structure*. Amsterdam: John Benjamins, vol. 28, pp. 23–41.
CourtListener. 2024. Available online: https://www.courtlistener.com/ (accessed on 6 December 2020).
Douglas, Charles G. 1983. How to Write a Concise Opinion. *Judges Journal* 22: 4.
Federal Judicial Center. 2013. *Judicial Writing Manual, A Pocket Guide for Judges*, 2nd ed. Washington, DC: Federal Judicial Center.
Federal Judicial Center. 2020. *Law Clerk Handbook*, 4th ed. Traverse City: Independently Published.
Fiddler, Garrett. 2020. SCOTUS Opinions. Available online: https://www.kaggle.com/datasets/gqfiddler/scotus-opinions (accessed on 10 November 2023).
Gozdz-Roszkowski, Stanislaw. 2020. Move Analysis of Legal Justifications in Constitutional Tribunal Judgments in Poland: What They Share and What They Do Not. *International Journal for the Semiotics of Law Revue Internationale de Sémiotique Juridique* 33: 581–600. [CrossRef]
Hafner, Christoph A. 2014. Professional Communication in the Legal Domain. In *The Routledge Handbook of Language and Professional Communication*. London: Routledge, pp. 349–62. Available online: https://api.taylorfrancis.com/content/chapters/edit/download?identifierName=doi&identifierValue=10.4324/9781315851686-29&type=chapterpdf (accessed on 10 November 2023).
Hartig, Alissa J., and Xiaofei Lu. 2014. Plain English and Legal Writing: Comparing Expert and Novice Writers. *English for Specific Purposes* 33: 87–96. [CrossRef]
Hiltunen, Risto. 2012. The Grammar And Structure Of Legal Texts. In *The Oxford Handbook of Language and Law*. Oxford: Oxford University Press. [CrossRef]
Hyland, Ken. 2012. ESP and Writing. In *The Handbook of English for Specific Purposes*, 1st ed. Edited by Brian Paltridge and Sue Starfield. Hoboken: Wiley, pp. 95–113. [CrossRef]
Kahn, Paul W. 2016. *Making the Case: The Art of the Judicial Opinion*. New Haven: Yale University Press.
Kalamkar, Prathamesh, Aman Tiwari, Astha Agarwal, Saurabh Karn, Smita Gupta, Vivek Raghavan, and Ashutosh Modi. 2022. Corpus for Automatic Structuring of Legal Documents. *arXiv* arXiv:2201.13125.
Klein, Richard B. 1995. OPINION WRITING ASSISTANCE INVOLVING LAW CLERKS: WHAT I TELL THEM. *Judges' Journal* 34: 33–36.
Leubsdorf, John. 2002. The Structure of Judicial Opinions. *Minnesota Law Review* 86: 447–95.
Lu, Xiaofei, Jungwan Yoon, and Olesya Kisselev. 2021. Matching Phrase-Frames to Rhetorical Moves in Social Science Research Article Introductions. *English for Specific Purposes* 61: 63–83. [CrossRef]
Maley, Yon. 1985. Judicial Discourse: The Case of the Legal Judgment. *Beiträge Zur Phonetik Und Linguistik* 48: 159–73.
Martin, James R. 2009. Genre and Language Learning: A Social Semiotic Perspective. *Linguistics and Education* 20: 10–21. [CrossRef]
Maswana, Sayako, Toshiyuki Kanamaru, and Akira Tajino. 2015. Move Analysis of Research Articles across Five Engineering Fields: What They Share and What They Do Not. *Ampersand* 2: 1–11. [CrossRef]
McKinney, Ruth Ann. 2014. *Reading like a Lawyer*, 2nd ed. Durham: Carolina Academic Press.
Miller, Carolyn R. 1984. Genre as Social Action. *Quarterly Journal of Speech* 70: 151–67. [CrossRef]
Moreno, Ana, and John Swales. 2018. Strengthening Move Analysis Methodology towards Bridging the Function-Form Gap. *English for Specific Purposes* 50: 40–63. [CrossRef]
Norman, Guy J. 2003. Consistent Naming in Scientific Writing: Sound Advice or Shibboleth? *English for Specific Purposes* 22: 113–30. [CrossRef]
Rampin, Rémi, and Vicky Rampin. 2018. Taguette: Open-Source Qualitative Data Analysis. *Journal of Open Source Software* 6: 3522. [CrossRef]
Sheppard, Jennifer. 2008. The 'Write' Way: A Judicial Clerk's Guide to Writing for the Court. *University of Baltimore Law Review* 38: 73.
Swales, John. 1990. *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge University Press.
Swales, John. 2004. *Research Genres: Explorations and Applications*. Cambridge Applied Linguistics. Cambridge: Cambridge University Press. [CrossRef]
Swales, John. 2016. Reflections on the Concept of Discourse Community. *La Revue Du GERAS ASP* 69: 7–19. [CrossRef]

Tarone, Elaine, Sharon Dwyer, Susan Gillette, and Vincent Icke. 1998. On the Use of the Passive and Active Voice in Astrophysics Journal Papers: With Extensions to Other Languages and Other Fields. *English for Specific Purposes* 17: 113–32. [CrossRef]
Tribble, Christopher. 2009. Writing Academic English—A Survey Review of Current Published Resources. *ELT Journal* 63: 400–417. [CrossRef]
Tribble, Christopher. 2015. Writing Academic English Further along the Road. What Is Happening Now in EAP Writing Instruction? *Elt Journal* 69: 442–62. [CrossRef]
van Geel, Tyll. 2009. *Understanding Supreme Court Opinions*, 6th ed. London and New York: Routledge, Taylor & Francis Group.
Vance, Ruth C. 2011. Judicial Opinion Writing: An Annotated Bibliography. *Legal Writing Inst.* 17: 197.
Yang, Yiying, and Fan Pan. 2023. Informal Features in English Academic Writing: Mismatch between Prescriptive Advice and Actual Practice. *Southern African Linguistics and Applied Language Studies* 41: 102–19. [CrossRef]

languages

MDPI

Article

Public Discourse on Criminal Responsibility and Its Impact on Political-Legal Decisions: Analysing the (Re-)Appropriation of the Language of Law in the Sarah Halimi Case

Nadia Makouar [1,2]

[1] Praxiling, Université de Montpellier Paul Valéry, 34090 Montpellier, France; nadia.makouar@univ-montp3.fr
[2] Aston Institute for Forensic Linguistics, Aston University, Birmingham B4 7DU, UK

Abstract: This applied linguistics study on the lay discourse about legal language analyses online public reactions to a court decision in the Sarah Halimi case, a French Jewish woman killed by her neighbour in Paris in 2017. This study draws on discourse analysis with a focus on semantics analysis and dialogism theory to delve into how legal discourse is disseminated in forums and undergoes semantic redefinition through users' language practices of legal notion in their own discourses. Thus, the aim of this study is not to develop linguistics theories but to use linguistics to explore the relationship between (1) the public representation and perception of this murder case in three forums and (2) the politico-legal response to decisions about a lack of criminal responsibility. The latter remains a sensitive topic in several countries, and several criminal justice reforms are revised or implemented with close observation of public reaction. This analysis highlights the linguistic markers revealing emotional discourse and a polymorphous expression of a lack of confidence in the justice system and legal actors, emphasising issues in comprehending justice and the work of psychiatrists and highlighting a gap between expectations and the actual delivery of justice. This study also shows that the linguistic strategies of non-experts are similar to those of legal experts.

Keywords: applied linguistics; language and law; criminal responsibility; penal populism; lay discourse; online discourses; semantics; dialogism

Citation: Makouar, Nadia. 2024. Public Discourse on Criminal Responsibility and Its Impact on Political-Legal Decisions: Analysing the (Re-)Appropriation of the Language of Law in the Sarah Halimi Case. *Languages* 9: 313. https://doi.org/10.3390/languages9100313

Academic Editor: Jeanine Treffers-Daller

Received: 8 December 2023
Revised: 22 July 2024
Accepted: 12 September 2024
Published: 27 September 2024

1. Introduction

The circulation of legal discourse online has been the subject of extensive research, with many studies examining these interactions and identifying a range of linguistic strategies. However, most studies in this field have focused on the communication of legal experts to non-specialists seeking legal advice or explanations (Diani 2023; Anesa 2016; Turnbull 2018a, 2018b). To illustrate this, Diani (2023) analysed the dissemination of knowledge in English and Italian forums, with a particular focus on the utilisation of explanatory structures, including denominations, definitions, descriptions, reformulations, paraphrases, exemplifications and generalisations. In another study, Diani (2022) explored the discourse on blogs specialising in law and the comments on posts. From a methodological point of view, the study involved two key approaches: (1) contrastive and qualitative analysis to compare posts and comments and (2) a qualitative study of a "dialogic action game", which they defined as "looking at blog posts and comments in terms of their speech acts and their initiative and reactive function" (Diani 2022, p. 11).

In the context of research analysing discourse in online forums with non-expert users, Demonceaux (2022) undertook an analysis of the dynamics of digital exchanges around the topic of homeopathy. The author identified a trend towards a "horizontalisation of discourse", which enables non-experts to share their experiences with controversial subjects. They can engage in a more egalitarian or symmetrical mode of communication. The author additionally observed that "health discussion forums [. . .] reflect a less vertical vision of health, opposing 'the normativity of medical discourse, where knowledge is transmitted

unilaterally from health professionals and medicines to the general public'" (Demonceaux 2022, online).

To the best of our knowledge, no studies have explored the linguistic and communication strategies of non-experts when speaking about legal notions yet. This paper attempts to contribute to the existing political science and law literature on penal populism by analysing the linguistic strategies deployed by non-experts in a non-expert online forum. The following was mentioned by Pratt (2007, p. 12):

> "Penal populism speaks to the way in which criminals and prisoners are thought to have been favoured at the expense of crime victims in particular and the law-abiding public in general. It feeds on expressions of anger, disenchantment and disillusionment with the criminal justice establishment. It holds this responsible for what seems to have been the insidious inversion of commonsensical priorities: protecting the well-being and security of law-abiding 'ordinary people', punishing those whose crimes jeopardize this."

Consequently, emotions predominate over reason, as penal populism depends on and fosters a fear of crime, portraying it as an escalating threat to society, faults the justice system and its purported ineffectiveness and advocates for more severe punishments and stringent measures against those who commit crimes (Boda et al. 2015). Thus, penal populism raises questions about the intelligibility of the legal system. In a study investigating media and sentencing within the French context, Philippe and Ouss (2016, online) conducted research on the influence of French media, specifically examining how television broadcasts of criminal justice events affect sentencing. The study showed that the duration of sentences extends by three months when the verdict is delivered subsequent to crime coverage. The lengthening of sentences is linked to the media's attention to the crime rather than the crime itself, and this impact diminishes rapidly. In this study, they demonstrated the impact of news content on criminal justice decisions. Their findings revealed that sentences in jury trials tend to be extended when there is increased coverage of crimes, while they tend to be shortened after the reporting of judicial errors. It is noteworthy that only media coverage related to crime and criminal justice, as opposed to coverage of other distressing subjects, exerts an influence on sentences. Additionally, their research showed that the timing of media coverage is crucial, with sentences being affected only by the reporting of crimes on the day immediately preceding sentencing rather than on other days. In contrast, they observed no discernible effect of media coverage on the sentencing decisions made by professional judges. This highlights the pivotal influence wielded by the media on laypeople's court decisions. Moreover, it underscores the susceptibility of lay jurors to such influence.

The decision to focus on non-specialist forums was made to gain insight into the circulation of legal concepts related to lack of criminal responsibility within this particular discursive space and to examine the ways in which they are received and discussed on these forums.

The discourse of non-experts is observed in two types of forums: general discussion in a video games forum with a high level of popularity among a young community (Gauducheau and Michel 2023; Durand 2017) and others which are more focused on politics or debates.

This article conducts a semantic analysis of forum discussions related to the Sarah Halimi case. Its objective is to explore how the public engages with legal language in their own discourses in online forums, particularly in the aftermath of the controversial decision in the case. This study intends to apply linguistics to analyse the relationship between public representation and perception of murder cases and politico-legal responses regarding diminished responsibility. The latter remains a sensitive topic in several countries, and a number of criminal justice reforms are revised or implemented with close observation of public reactions and perceptions of justice (Noyon et al. 2020).

2. Context

Sarah Halimi was a French Jewish woman killed by her neighbour, Kobili Traoré (K.T.), in Paris in April 2017. The psychiatric assessments concluded that the murderer committed the crime during an "acute delirious puff" against a background of heavy cannabis consumption. His defence argued that he was suffering from a drug-induced psychotic episode at the time of the murder. According to expert reports, Kobili Traoré was suffering from an acute delirium linked to heavy cannabis consumption. Psychiatric assessments led the examining magistrate's chamber to conclude that he had lost his discretion. The expert reports and counter-expertise clashed, creating a controversial situation in public opinion and among certain politicians in 2019.

The Court of Cassation (the highest court in the French judiciary system), while confirming the antisemitic nature of the crime, maintained the lack of criminal responsibility of the murderer. This decision sparked strong reactions in France and worldwide during the next few weeks and significant engagement on social media. Faced with these reactions, the public prosecutor (Magistrate François Molins) admitted that "the emotion aroused by this decision probably reveals that the current law is not appropriate".

On the side of legal professionals, there has been a call for elucidation of the legal framework regarding criminal liability in the event of the voluntary consumption of psychoactive substances. This has led to a suggestion that the parliament should address the ambiguity of the law. At the same time, following these claims, both from public opinion and from legal professionals, the Minister of Justice announced in April 2021 a draft law on lack of criminal responsibility. This bill aims to "fill" a "legal vacuum," which sparked debates within the legal sphere[1].

Public opinion played a significant role in the political and judicial consequences of this case. The issue of a lack of criminal responsibility generated a sense of injustice within public sentiment. On the legal side, concerns arose regarding disruptions to the judicial system, prompting the prosecutor to emphasise the importance of maintaining strict independence. Additionally, it was stressed that any changes to the law should be approached with caution and not implemented "in a hurry and in the heat of the moment"[2]. However, public opinion is explicitly considered in political-legal texts. In fact, on 26 May 2021, the Conseil d'état (the highest administrative court) considered the draft law limiting lack of criminal responsibility. Within this text, public opinion is recognised as one of the aspects addressed in the bill and explicitly acknowledged by the Council of State[3]:

> "[. . .] However, [the Conseil d'état] stresses that the exception introduced by the draft law, which is intended to respond to the emotion aroused in public opinion by tragic events, is more than limited in scope, as the conditions for exclusion from lack of criminal responsibility appear to be very theoretical and proof of the intentional element extremely difficult to provide in practice. [. . .] Some of the provisions of the bill—those relating to lack of criminal responsibility or the creation of a new offence to punish certain acts of violence committed against members of the gendarmerie and police officers in particular—were decided by the government following events that aroused great emotion in public opinion." (Conseil d'etat, Avis consultatif 2021)

These discourses surrounding the two court decisions in 2019 and 2021 show the extent to which the case has aroused much emotion, many reactions and the need to explain and change the law.

2.1. Judicial Chronology of the Case

In December 2019, K. T. was declared not criminally responsible. This decision provoked strong reactions because many believed that antisemitic motives played a significant role in the crime. The case was then referred to the Court of Cassation to push for a trial in a criminal court.

In April 2021, the Court of Cassation validated the previous decisions and considered that the provisions of the current law "do not distinguish based on the origin of the mental disorder that led to the abolition of this discernment [. . .][4]". Thus, the court confirmed the declaration of lack of criminal responsibility and upheld the antisemitic nature of the crime. K. T. was hospitalised in a psychiatric unit.

2.2. Criminal Responsibility and Public Opinion

According to Fovet et al. (2022), criminal responsibility "has been a core principle of French criminal law since the early nineteenth century". This notion stands as a pivotal concept within the criminal sanctions applied to individuals diagnosed with mental health disorders.

A few political and legal figures made metatextual comments in the media following this court decision. Political reactions about "voluntary intoxication" came from all sides, but they also came from politicians, including President Macron.

In April 2021, after the court confirmed the lack of criminal responsibility of K.T., E. Macron commented on the decision:

> *In my opinion, deciding to take drugs and then becoming "mad" should not remove your criminal responsibility. I would like the Minister of Justice to propose a change in the law as soon as possible.*[5] *(E. Macron, Le Figaro, April 2021)*

From the same perspective, and at the same time, the Minister of Justice, Eric Dupont-Moretti, initiated proposals to fill a legal vacuum regarding this decision.

Macron's statement, along with that of Justice Minister Eric Dupont-Moretti, showed the impact of public opinion on political decisions.

From the legal side, Prosecutor Molins, who had previously expressed concerns alongside the court president, Chantal Arens, regarding politicians' comments on the 2019 Court of Appeal decision in an official statement[6], asserted that justice has fulfilled its role. The legal proceedings have acknowledged the commission of an antisemitic crime, but on the grounds of lack of criminal responsibility, this does not grant the court the authority to prosecute K.T.

2.3. "Legal Vacuum" and Law Changes

As specified above, a few days after the decision, the president announced that he wanted a law ruling out lack of criminal responsibility on the grounds of psychic or neuropsychic disorders in cases of drug use. The political leaders in France have considered that the law should be changed for several points. The law proposed in an accelerated procedure in 2021 and voted on in January 2022 limits lack of criminal responsibility in cases of mental disorders resulting from voluntary intoxication with psychoactive substances. It includes the following measures (Clément 2022):

- Exclusion of lack of criminal responsibility in the case of voluntary intoxication;
- Exclusion of reduced criminal liability in cases of voluntary intoxication;
- Creation of voluntary intoxication offences.

For example, for the measure related to the exclusion of reduced criminal liability in cases of voluntary intoxication, the change occurs in the second paragraph of article 122-1 of the Criminal Code. It "reduces the penalty incurred by a person whose discernment or control of his or her actions has been impaired, but not abolished, by a psychic or neuropsychic disorder". Clément (2022) noted that the new article (122-1-2) of the same code will similarly exclude from this reduced penalty anyone who voluntarily, illicitly or manifestly excessively consumes psychoactive substances.

3. Theoretical Framework and Data

3.1. Research Questions and Data

Research questions have arisen to understand the discourse developed between media discourse (referenced in the forum posts) and that of court decisions. The current project

involves comprehending how the discourse from the latter, identified as a specialised form of communication, is reformulated or used in the discussions of internet users who lack legal expertise.

Thus, the research questions are the following: (1) How do users on forums comment on this case? (2) What are the perceptions of various actors, including lawyers, judges and experts? (3) How do they repurpose legal language in their discussions? How can this analysis help with understanding non-expert discourses, their skepticism and the lack of intelligibility?

The legal and political context following this highly controversial decision led me to question the public discourse. I chose to analyse the discourse of internet users on three forums which dealt with the case extensively in 2019 and 2021. The aim was to investigate the lay discourse of the language of law, especially the processes of reformulating decisions and legal facts, and analyse the development of opinions about what should have been decided. Without claiming to be exhaustive or representative of online discourse on this issue, I chose three forums (Accessed on June 2023).

Jeux-videos is a website originally dedicated to video games, but it has broadened its focus to include debates on politics and discussions of everyday problems (Lamy 2017). Regarding *Forum-Actualite,* this is a discussion forum open to "debates on politics and sport". As for *the Forum Politique*, it is presented as "a French-speaking forum dealing with political and social issues in general. Its aim is to enable contributors to discuss all the subjects indicated by the forum's sub-headings, to exchange information, and to compare ideas" (forum-politique.fr). The corpus consists of 707 posts (see Table 1). All posts have been translated by the author of this paper from French to English. The original posts are available in the Appendix A.

Table 1. Forums' corpora.

Forums	Posts	Dates
Jeux-videos (JV1)	34	December 2019
Jeux-videos (JV2)	233	April 2021
Forum-actualité (FA)	173	June 2017–May 2021
Forum-politique (FP)	267	December 2019–April 2023

3.2. Theoretical Framework and Methods

This study draws on discourse analysis with a focus on semantics analysis and dialogism theory to delve into how legal discourse is disseminated in forums and undergoes semantic redefinition through users' language practices of legal notion in their own discourses. The objective of this applied linguistics study is to comprehend the circulation of legal discourse on forums and how the semantic reinterpretation of legal terms by internet users unfolds. Additionally, this study seeks to identify this phenomenon by examining the conditions of intertextuality, specifically the reuse of discourses in different spaces. Indeed, according to Garric and Longhi (2013, p. 65):

> "Discourses do not belong to delimited zones of practice. Situated in interdiscourse, considered a dynamic and conflictual space of circulation, they are traversed and invested by social objects that take on meaning in the plurality of interpretative paths in which the subject participates by assuming different successive sociodiscursive roles."

Thus, the circulation of discourses has an impact on the characteristics they take on according to the modes of transmission. According to Longhi and Sarfati (2007), these "can also give rise to argumentative manipulations", and a "discourse can subvert the characteristics of another type of discourse in order to take advantage of its specific characteristics" (Garric and Longhi 2013).

According to Rastier (2011), the semantic interpretation of words is only possible because the terms are adjacent to each other. This means that the same term, however specific, can have a different meaning when present in the text of a different genre. Rastier also argued that any text placed in a corpus receives semantic determinations and can potentially modify the meaning of each of the texts which comprise it.

The interaction between discursive universes involves semantic redefinition phenomena due to the conditions under which the meanings of terms are interpreted. Depending on the textual genre and social practices, the conditions of interpretation are reconfigured, particularly in terms of enunciative foci. According to Rastier (2017), this is related to the fact that the norms of discourse, genre, and style are anchored in social practices. Also, these norms "bear witness to the impact of social practices on the texts they govern" (Rastier 2017, p. 12).

In the example of our data, we assumed that legal citations or terms, as soon as they were transposed into forum posts, received other semantic determinations. This means that a legal term A will have semantic features specific to the legal genre text in which it is expressed. If this term A is transposed by a quote, such as in a forum post, then it is likely to receive different semantic features and therefore have a different meaning due to the context of the utterance.

In interpretive semantics, the *seme* (or *semantic feature*) is the smallest unit of meaning. Two types of *semes* are identified: inherent *semes* and afferent contextual *semes*. Rastier and Riemer (2015, p. 494) defined an "inherent *seme*" as an attribute having a typical value. On the other hand, an "afferent contextual *seme*" is one that is activated by the linguistic context. The meaning of a word can be "perceived" with interpretative operations such as activation, inhibition or propagation (Rastier and Riemer 2015, p. 495).

The notion of dialogism is the second aspect which completes the theoretical framework on which this study is based. The concept of dialogism finds its origins in the scholarly contributions of the cercle de Bakhtine. According to Brès (2017), "dialogism thus consists in the orientation of any discourse (whatever its format: speech, press article, political discourse, scientific article, literary text, etc.) towards other discourses in the form of an internal dialogue with them." This perspective of dialogism makes it possible to take account of the multiple voices which an utterance may contain by considering the point of view of the enunciator and the different speakers who are quoted or taken up in the forums. This perspective on dialogism, particularly the interdiscursive dialogism aspect, is significant in understanding the dynamics of discourse and communication.

Theoretical research on penal populism is also one of the foundations of this study. In that perspective, if penal populism is associated with disinformation and conspiracy theories, then the challenge of this analysis in legal linguistics is also to identify the discursive mechanisms revealing mistrust of institutions, including conspiracy discourse in relation to the political and legal systems. Many research studies have demonstrated that discussion forums provide a conducive environment for the belief, emergence and dissemination of conspiracy theories and disinformation (Shahsavari et al. 2020; Allington et al. 2021). It is also a space where individuals discuss the credibility of conspiracy theories on online forums (Bangerter et al. 2020). Also, according to Douglas et al. (2019), some conspiracy theories can satisfy important social psychological motives. These motivations can be epistemic (e.g., the desire for understanding, accuracy and subjective certainty), existential (e.g., the desire for control and security) and social (e.g., the desire to maintain a positive image of the self or group). The objective is to highlight through discourse analysis the issue of the intelligibility of the legal and legislative systems and to understand the discursive strategies used by internet users to explain, understand or express their approval or disapproval of legal decisions. This holds significant importance for legal linguistics insofar as linguistic analysis may reveal a semantic discontinuity and a missing interpretative link between media discourse, which is a primary source for the public in the forums I analysed, and the legal discourse in court decisions.

The discussions on online forums involve participants from diverse social backgrounds, exhibiting a wide range of expertise and knowledge levels, and researchers in linguistics, including socio-terminologists, encourage adopting a scalar approach rather than a binary one when examining discourse, moving away from a strict division between specialised and lay discourse (Gaudin 2003; Vicari 2018).

Analysing public debates helps to reveal linguistic strategies which evoke emotional responses and support for punitive measures. Hence, describing and identifying the way certain terms and phrases relate to a penal populism narrative and how it influences legislative and legal discourses and practices is crucial for legal linguistics. In sum, this study seeks to identify the discursive and metadiscursive mechanisms of penal populism in lay discussions within online forums, specifically focusing on how individuals express their views on legal discourse by using it. The objective is to examine patterns in the way internet users perceive judicial and political institutions, as well as legal concepts like "lack of criminal responsibility", through linguistic and semantic analysis.

Based on this theoretical framework, the analysis method consisted of identifying recurring themes running through the three forums. To accomplish this, I followed the following steps:

(1) First, observe the way in which topics are created and initiated. The objective is to understand the motivations of these posts.
(2) Then, identify the cooccurrences of "Sarah Halimi", "Kobili Traoré" and "justice", "penal responsibility" or "irresponsibility" to understand how these legal notions are reformulated, defined and qualified.
(3) Finally, identify the cooccurrences of the various legal actors to comprehend the perceptions of the legal and political actors.

4. Findings

What was noticed quite immediately was that topics were mostly initiated by a reference to a media article. The internet user opens the forum topic by providing key information and making a comment. The sequence of interactions is based on the title or content of the article and the legal nature of the case.

4.1. Initiating Topics with Media Articles Related to the Court Decision

In the three subjects initiated in December 2019 and April 2021, the dates on which the courts handed down their judgements, forum users began by publishing the headline and the link to the article in all of the topics selected for the analysis. On an interdiscursive level, legal discourse is transposed first into media discourse and then into forums. In December 2019, the Court of Appeal confirmed K.T.'s lack of criminal responsibility and his hospitalisation in a psychiatric unit. After the civil parties appealed to the Court of Cassation, the latter confirmed the previous decisions in April 2021:

(1) The murder of Sarah Halimi

> *"We have just created the Sarah Halimi jurisprudence in our country, meaning that anyone who suffers a delirious episode because they have taken an illegal substance that is dangerous to their health will be exonerated from criminal liability", he warned. [link to a media article] (FP, 19 December 2019)*

In example (1), the commenter used a quote from the family's lawyer (Mr. Szpiner), which they highlighted in the body of the topic as an authoritative argument to emphasise the indignation it aroused by using the term "jurisprudence". This term's use underlines the unprecedented and exceptional nature of the decision which was handed down. It is mainly this discourse which internet users understand and use to initiate debate.

In example (2), the user repeats the first sentences of the article. In example (3), the user explains his understanding of the article:

(2) Sarah Halimi's murderer won't be judged.

On Thursday, Kobili Traoré was declared not criminally responsible at the time of the events in 2017. On Thursday, December 19, the Paris Court of Appeal ruled that Sarah Halimi's murderer was not criminally responsible for the events of 2017, as reported by Le Figaro. [. . .] [link to a media article] (JV1, 19 December 2019)

(3) No trial for SARAH HALIMI: Justice has decided

No trial for Sarah Halimi, killed by Kobili Traoré who is considered not criminally responsible. The judges applied the law. Simply. [link to a media article] (JV2, 15 April 2021)

The identification of dialogical phenomena and the circulation of media discourse within forums is crucial because the media serve as the primary information source and hold authoritative sway in forums, initiating the debate. This underscores the significant impact of the media and their discourse on internet users.

4.2. Discussing Laws and Legislative Texts

4.2.1. Discussing Lack of Criminal Responsibility and the Judge's Decision

It is noticeable that in some posts, users defended the judge's position and tried to explain the points made in the article. This is the case in example (4) below:

(4) "The judge cannot distinguish what the legislator has chosen not to distinguish."

In other words, it is the legislature's fault for having framed the criminal law in question too narrowly, and the judges are therefore inviting the legislature to adopt a new law along these lines.

"The judge is the moth of the law," said Montesquieu, and we have a perfect illustration of that with this ruling. (JV2, 15 April 2021)

This example shows the epistemic stance of the forum user, who reformulated the comments for a less specialised audience. As stated by Hyland (2007, pp. 268–69), "Reformulation is a discourse function whereby the second unit is a restatement or elaboration of the first in different words, to present it from a different point of view and to reinforce the message." The user posted part of the article on the role of the judge and legislation (quotations markers) and put themself in a position to explain to future respondents how the judicial process and legal decisions work, opening their remarks with the meta-discursive marker "in other words" and adding a philosophical reference to support their argument.

The discussion then turned to the legitimacy and responsibility of taking drugs. In example (5), the author of the post insists on the voluntary nature of taking drugs. In their view, there is no reason to remove responsibility from the murderer:

(5) you take drugs with KNOWINGLY, the famous abolished discernment, whereas he follows a religious logic in his crime, we are more on a total disinhibition than an abolishment of the discernment (JV2, 15 April 2021)

In this example, the user sought to change the wording by using "disinhibition". In the same sentence, the adjective "famous" is used to express irony regarding the legal notion and opposes it with religious logic, being associated with a supposed planned crime. This reasoning allows them to assert that the murderer is responsible for his crime and that drug used allowed him to disinhibit himself. Thus, the semantic of responsibility is activated.

4.2.2. Analogy with Alcohol Consumption

Several posts in the four subcorpora referred to alcohol consumption. Forum users used this to understand and express their views on the reasoning of the courts and the law regarding lack of criminal responsibility.

In example (6), the user argues that alcohol was considered a mitigating circumstance. According to their understanding, the use of drugs is also a mitigating circumstance here, just as the use of alcohol should become an aggravating circumstance:

(6) *It reminds me that drinking and driving used to be an extenuating circumstance when you had a serious accident, then it became an aggravating circumstance, and everyone thought it was crazy when they found out. [. . .] (JV1, 20 December 2019)*

(7) *(ba) no, you take drugs knowingly, so you're responsible for what you do under the influence. As far as I know, when you do something stupid and drunk, we punish you anyway. (hein) (JV2, 15 April 2021)*

In this example, the users mentioned shared knowledge about alcohol being an aggravating circumstance. "It reminds me" and "as far as I know" introduce this shared knowledge. In example (7), the use of "ba" and "hein", which are locutions referring to something obvious and commonly known, implies the epistemic stance of the user.

4.2.3. Imaginary Scenarios

In example (8), the user imagines an expeditious and lenient trial for the defendant, who had "8.6 g of alcohol and six joints in his brain". The defendant was therefore "acquitted". The author sums up their point of view with this imaginary trial, and the analogy here shows the simplification of the author's argument regarding K.T.'s lack of criminal responsibility. Even if it is inaccurate to say that K.T. had been acquitted, what is being pointed out here is that the decision is considered to be lenient towards the use of substances which lead to serious offences:

(8) *-You are accused of killing twelve pedestrians by running them over with your car. How do you plead?*
-I had 8.6 grams in my blood and six joints in my brain, your honour.
-Acquitted (JV2, 15 April 2021)

They contrast the seriousness of the circumstances ("8.6 g in my blood and six joints in my brain"; "traffic offender kills") with the supposed leniency of the sentence handed down to the perpetrator ("Acquitted"; "claim they are not responsible"; "trick is done"):

(9) *Tomorrow, when a traffic offender kills one person or multiple people and is driving under the influence of drugs or alcohol, they will be able to claim that they are not responsible because they were not themselves' at the time of the accident and were driving unconsciously. That's it, the trick is done, and already the 'justice system' is decriminalising drug use by deeming that the person who has taken drugs is not responsible, either for their consumption or for what they do afterwards! (FP, 19 December 2019)*

These scenarios were used by the forum users to express their disagreement with the decision. This strategy was also used in lay-legal interaction with a different goal. The work of Diani (2023, p. 305) defines, "Scenario, which consists in illustrating possible or hypothetical situations, more complex events, or reactions, and taking into consideration a broader context, to refer to the specific situation".

The user in example (9) comments about the decriminalisation (*décriminalisation*) of drug use in relation to lack of criminal responsibility. In law, decriminalisation "means that the legislator passes a law stating that an illegal act will no longer be an offence in the future. In other words, prohibited behavior is transformed into permitted behavior[7]". By using the term "decriminalise", the user extrapolates the decision of the Cour de Cassation, claiming that it is a question of decriminalising drug use. This argument assigns the semantic features of laxism and permissive to the term "justice".

4.2.4. Irony and Conspiracy

There are several posts with ironic content in the JV1 and JV2 forums. Irony is an argumentative process and "can be considered a pivotal strategy, positioned somewhere between discourse destruction and refutation. Irony ridicules a speech that pretends to be dominant or hegemonic, by implicitly referring to some contextually available irrefutable rebutting evidence" (Plantin 2021, online).

In examples (10) and (11), the authors use irony to simplify the facts and disapprove of the decision. This is even more obvious in example (12), stating "breaking the law [...] nullifies the crime", which serves to emphasise the paradox of the decision:

(10) So he had the right to murder her after taking pot (JV2, 15 April 2021)
Answer to (10)

(11) yes, breaking the law by taking drugs and then killing someone under the influence of said drugs cancels out the crime (JV2, 15 April 2021)

Conspiracy theories also featured prominently in the four subcorpora. The examples below show that the author thinks that K.T. is being favoured because he is Muslim. Also, in comment (12), the user uses "you-know-who" for designation but without naming who or what is involved, though this was possibly understood by other users. This is also a strategy to prevent the message from being deleted by moderators:

(12) There's been a lot of this "criminally irresponsible" stuff lately. It's the new term for you-know-who (JV1, 19 December 2019)
(13) What's becoming very alarming in France is that more and more criminals are being judged irresponsible—all they have to do is shout "allah what's-his-name" and that's it, you're mentally deficient (which, in a way, is not wrong) (FA, 20 December 2019)
(14) [...] The same firm that defended the leftist Cedric Herou, who smuggled illegal immigrants in defiance of the law [. . .] Politicized justice system. Progressive, pro-immigration, left-wing lawyer (obviously) [...] (FA, 19 April 2021)

In example (13), there is a similar allusion to the fact that clemency is granted to Muslims. According to the user, "they have to shout "*Allah whatever*" and that's it, you're mentally deficient". In example (14), the user comments that lawyers are manipulated, and everything is then allowed, evoking the idea of laxism.

These elements come close to conspiracy theories about immigration and the complicity of the left (Makouar 2022) because of Kobili Traoré's foreign background and the anti-Semitic motive.

4.3. Discourse on the Perception of Justice and Psychiatric Experts

The perception of justice is an element which runs through our four subcorpora. The justice system and those involved in it are sometimes viewed negatively. In example (15), psychiatrists are described as "sick" and unprofessional because they base their decisions on their "thoughts" and not on facts:

(15) Justice in France is becoming increasingly ridiculous. Well, there are sick people called psychiatrists who spout off their 'analyses' based on what they think they know and the judges who go along with it. (FP, 19 December 2019)

In example (16), the user has a similar opinion, saying that too much confidence is placed in psychiatric assessments and implicitly comparing psychiatric hospitals and prisons. However, some comments put these opinions into perspective (17). The user explains that the French justice system is poor in law and financially ("It applies *badly* voted laws") and does what it can with the resources at hand:

(16) That's the truth, unfortunately. In my opinion, too much importance has been given to psychiatric assessments. The guy will probably spend the rest of his life in a psych ward, but it's still a bit disgusting (JV2, 15 April 2021)
(17) The French justice system does its job as best it can.
It applies badly voted laws, being the poorest justice system in Europe (JV1).

4.4. Discussing the Perception of Psychiatric Hospitals and Prison Environments

As previously mentioned, the notion of incarceration is strongly associated with the idea of justice. Without incarceration in prison, even if the murderer must undergo several years of psychiatric care, prison remains the only solution which ensures a sense of security

and justice. This narrative suggests that psychiatric hospitals are places of "leisure". This can be observed in example (18), where it is associated with "Club Med[8]":

(18) You've watched too much "One Flew Over the Cuckoo's Nest", and the drug overdoses and lobotomies are over. The psych ward has become a branch of Club Med. (FP, 19 December 2019)

Thus, the semantic features of pleasure and wellness are propagated to prison, suggesting there are comfortable places and activating a resentment of injustice. The comments also suggest that confinement and punishment should be definitive, with no way back. This can be observed in example (19), where the internet user associates bad psychiatric assessments with a "tragedy" which happened before and argues that security for the society depends on the lifelong incarceration of K.T.:

(19) What you don't want to understand is that there are already precedents for releases validated by psychiatrists that have led to real tragedy, so in truth, we're mainly worried about the omnipotence of psychiatrists over this kind of decision. If I'm guaranteed that his condition is not compatible with life in society and that he's therefore locked up for life despite a possible recovery, then I'm all for it. (JV2, 15 April 2021)

(20) He will not be incarcerated but will be interned in a unit for dangerous patients for life and put on heavy treatment with no contact with normal individuals. Personally, I would prefer prison or death... (FA, 20 December 2019)

Also, some users tempered this narrative on punishment and tried to explain what happens when people are interned in psychiatric wards (20).

5. Discussion and Conclusions

Drawing on the theoretical framework of dialogism and interpretative semantics, this qualitative analysis of the corpus has highlighted the ways non-expert discourse attempts to understand, explain and refute the decision rendered, being directly linked to the question of lack of criminal responsibility. The linguistic mechanisms revealed that internet users have employed various strategies to express their opinions. Some of them are similar to explanatory structures used by legal experts in forums dedicated to legal advice, such as reformulation, denomination and scenarios (Anesa 2016; Diani 2023). This study also revealed that forum initiators primarily rely on media texts to open discussions, engaging in a dialogical process. The media served as a key information source within the forums analysed in this paper.

Research showed that court rulings frequently draw substantial media attention and are subject to diverse interpretations by internet users. In criminal cases, as exemplified by Salas (2021, online), public sentiment can be shaped by a punitive inclination, especially when mass media intensifies public anger while still maintaining an attachment to a humane approach to penalties. The confluence of a criminal news event, political discourse, mass media coverage and the absence of alternative perspectives tends to foster a punitive reaction. For Salas, this is where misinformation and the dissemination of preconceived notions find fertile ground, such as in claims that the death penalty can effectively reduce crime. This is what we observed in the corpus, in addition to the use of conspiratorial discourse. Thus, the impact of the media and a limited comprehension of the justice system by laypeople contribute to the emergence of penal populism. This phenomenon is marked by a discourse of mistrust of the justice and political systems, as well as the propagation of conspiracy theories suggesting that criminals receive special treatment to the detriment of victims.

Salas (2021, online) discussed a duality of media and legal scenes as well as a narrative war which intensifies, especially with the development and virality of social media. The topic and comments take on dimensions of controversial discourse and conspiracy. Indeed, a significant portion of the comments opposed the court's decision, using irony and reformulation, reusing discourses of legal professionals or politicians to change denomination and employing conspiracy discourses, imaginary scenarios or analogies to convey their understanding and refute the court's arguments. Discussions on forums revealed a lack of

confidence in the justice system, highlighting a gap between expectations and the actual delivery of justice. Internet users emphasised issues in comprehending justice and the work of psychiatrists.

What is particularly intriguing is the issue of the connection between the legal decision and the concept of "doing justice". There appeared to be a notable disparity between the expectations of public opinion and the perception of achieving justice for the victim. For many internet users, imprisonment was seen as the sole means to fulfill this objective. Justice was often conflated with punishment, prompting questions about the actual efficacy of imposed sentences.

As Pratt (2007, p. 173) pointed out, emotions and feelings of insecurity often take over when a criminal case breaks in the media. This ties in with Salas's (2021, online) analysis of public reaction and its relationship with the media. This issue is highly topical. Recently, in France, a criminal case (Affaire Lola) was widely reported in the media, which led the Minister of Justice Dupond-Moretti to intervene publicly. He voiced strong criticism of media coverage regarding this case on a French TV show (*Touche Pas à Mon Poste*, presented by Cyril Hanouna) and highlighted the temporal disconnect between the media and the justice system, saying that "[t]here is no room for populism when dealing with a tragedy like this [...] We must respect the rules that have taken thousands of years to develop"[9]. Thus, the issue of penal populism is becoming increasingly important, and we can assume that there is a significant degree of radicalism in the discourse on the punishment to be meted out.

By using semantics and dialogical theories, this study explored the political, legal, legislative and public opinion contexts of controversial court decisions. It is a first step towards a broader understanding of lay discourse, its engagement with public opinion and the role and influence which such discourse might play in the political and legal contexts. This study highlighted the linguistic features of lay discourse, demonstrating how the combination of linguistic approaches can identify the characteristics of non-expert discourse on the one hand and enhance the intelligibility of legal systems and discourse on the other. In other words, the applied approach and methods of linguistics could contribute to the development of more comprehensible discourse and provide key information for non-experts to understand legal issues and discourse.

The results may have several implications: pedagogical (law students), institutional or in the media (legal fact-checking). Salas (2021, online) argued that greater familiarity with the criminal justice system tends to correlate with reduced punitive attitudes among individuals. Furthermore, a deeper understanding of case particulars often leads views on punitiveness to align more closely with legal assessments. Through semantic analysis, the study identified discursive mechanisms associated with the difficulties in understanding legal language, the simplification of the legal system and its discourse and the sentiments of injustice linked to the concept of lack of criminal responsibility. Regarding this legal notion and its cooccurrences, the translation of the posts from French to English was not easy to achieve, especially for the term "irresponsabilité pénale", translated to "lack of criminal responsibility". Also, terms such as "jurisprudence" or "decriminalizing" were not easy to translate because of the differences of jurisdiction in different countries. The work in progress involves collecting media articles, tweets and forum posts during court rulings, aiming to identify phenomena of the circulation of legal texts or concepts to better understand the issues of the shared meaning of legal discourses when they circulate in other spheres and discursive genres. Methodologically, corpus linguistics can be employed to unveil perceptions of the justice system by analysing cooccurrences and understanding the extent to which discourse may be polarised for an issue.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The author declares no conflict of interest.

Appendix A. Original Comments in French (without Emojis)

(1) Meurtre de Sarah Halimi

"On vient de créer dans notre pays une jurisprudence Sarah Halimi, c'est à dire que toute personne qui sera atteinte d'une bouffée délirante parce qu'elle aura pris une substance illicite et dangereuse pour la santé se verra exonérée de responsabilité pénale", a-t-il mis en garde.

https://www.bfmtv.com/police-justice/meurtre-de-sarah-halimi-pas-de-proces-pour-le-suspect-juge-penalement-irresponsable-1827329.html (FP, posted on 19 December 2019)

(2) Le meurtrier de Sarah Halimi ne sera pas jugé

Kobili Traoré a été déclaré, ce jeudi, pénalement irresponsable au moment des faits, en 2017. Ce jeudi 19 décembre, la cour d'appel de Paris a jugé que le meurtrier de Sarah Halimi était pénalement irresponsable des faits, en 2017, rapporte notamment Le Figaro. [. . .]

https://www.valeursactuelles.com/faits-divers/le-meurtrier-de-sarah-halimi-ne-sera-pas-juge (JV1, posted on 19 December 2019)

(3) Pas de PROCÈS pour Sarah HALIMI: la JUSTICE a TRANCHÉ –

Pas de procès pour Sarah Halimi tué par Kobili Traoré considéré comme irresponsable pénalement. Les juges ont appliqué le droit. Tout simplement.

https://www.lemonde.fr/societe/article/2021/04/14/mort-de-sarah-halimi-la-cour-de-cassation-confirme-l-irresponsabilite-de-son-meurtrier-qui-ne-sera-pas-juge_6076764_3224.html (JV2, posted on 15 April 2021)

(4) "Le juge ne peut distinguer là où le législateur a choisi de ne pas distinguer".

Sous-entendu, c'est de la faute du pouvoir législatif d'avoir enfermé le texte pénal en question trop restrictivement et les juges invite donc ce dernier a adopté une nouvelle loi en ce sens.

Le juge est la bouche de la Loi disait Montesquieu et l'on a une illustration parfaite avec cet arrêt.

(5) tu consommes de la drogue en CONNAISSANCE DE CAUSE, le fameux discernement aboli alors qu'il suit une logique religieuse dans son crime on est plus sur une désinhibition totale qu'un abolissement du discernement (JV2, posted on 15 April 2021).

(6) Ça me fait penser qu'avant l'alcool au volant était une circonstance atténuante quand on avait un grave accident, puis c'est devenu une circonstance aggravante, and tout le monde trouve ça dingue quand il l'apprenne Là c'est pareil pour ce meurtre, sauf que y'a que la justice qui trouve ça normal de dédouané un meurtrier drogué au moment des faits, peut être qu'un jours la loi sera inversé comme pour l'alcool au volant (JV1, posted on 20 December 2019).

(7) ba non tu consomme de la drogue en connaissance de cause tu es donc responsable des actes que tu commets sous l'influence de la drogue. quand tu fais des connerie completement bourré a ce que je sache on te punit quand meme hein (JV2, posted on 15 April 2021).

(8) Vous êtes accusés d'avoir tués douze piétons en les écrasant avec votre voiture, que plaidez-vous ?
J'avais 8.6 grammes dans le sang et six joints dans le cerveau, votre honneur
Acquitté

(9) [. . .] Demain, lorsqu'un criminel de la route aura tué une ou plusieurs personnes, and qu'il avait pris le volant drogué ou alcoolisé, il pourra prétendre ne pas être respon-

sable car il n'était pas "lui même" au moment des faits et qu'il avait pris le volant inconsciemment. Voilà, le tour est fait, déjà la "justice" dépénalise la consommation de drogue en estimant que celui qui en a consommé n'est pas responsable, ni de sa consommation ni de ce qu'il fait après !

(10) Donc il avait le droit de l'assassiner après avoir consommé du shit (JV2, posted on 15 April 2021).

(11) oui enfreindre la loi en consommant de la drogue puis tuer quelqu'un sous l'influence de ladite drogue annule le crime (JV2, posted on 15 April 2021).

(12) ça commence à faire beaucoup ces derniers temps ces "pénalement irresponsable". C'est le nouveau terme pour vous savez qui ? (JV1, posted on 19 December 2019)

(13) ce qui devient très alarmant en France c'est que de plus en plus de criminels sont jugés irresponsables ils suffit qu'ils gueulent allah machin et ça y est ,tu es déficient mental (ce qui ,quelque part n'est pas faux) (FA, posted on 20 December 2019).

(14) Le même cabinet qui a défendu le gauchiste Cedric Herou, passeur de clandestins, au mépris de la loi (mais c'est GI qui a été dans le collimateur parce que ces jeunes voulaient faire respecter la loi). Justice pourave politisée. Avocat progressiste, pro immigration, de gauche (évidemment, sinon).

(15) La justice en France devient de plus en plus ridicule. Bon, il y a les malades appelés psychiatres qui débitent leurs "analyses" sur fondement de leur pensée de savoir et des juges qui marchent dans la combine. [. . .] (FP, posted on 19 December 2019).

(16) Exact c'est malheureusement la vérité, on a donné trop d'importance aux expertises psychiatriques à mon avis. Après le type passera sûrement sa vie en hôpital psy mais bon c'est quand même un peu dégoûtant (JV2, posted on 15 April 2021).

(17) La justice française fait son travail comme elle le peut. Elle applique des lois mal votées, en étant la justice la plus pauvre d'Europe (JV1, posted on 20 December 2019).

(18) Toi t'as trop regardé "vol au dessus d'un nid de coucou", les surdoses de came et les lobotomies c'est fini. L'asile de dingue c'est devenu une succursale du Club Med (FP, posted on 19 December 2019).

(19) [..] Ce que tu veut pas comprendre c'est que y'a déjà des précédents de libération validé par des psychiatres qui on mené à de véritables drame donc en vérité on s'inquiète surtout de la toute puissance de psy sur ce genre de décisions. . .moi si on me garantit que son état est pas compatible avec la vie en société et que du coup on l'interne à vie malgré une possible guérison alors je veut bien. [. . .] (1) (JV2, posted on 15 April 2021).

(20) Il ne sera pas incarcéré mais sera interné en unité de patients dangereux a vie, mis sous traitement lourd sans contact avec des individus normaux. personnellement, je préfèrerais la prison ou la mort... (FA, posted on 20 December 2019).

French-to-English translations in this article are by the author of this paper.

Notes

1 «Responsabilité pénale: l'ordre des avocats du barreau de paris s'oppose à un projet de loi fourre-tout, bâti dans la précipitation», 20 October 2021, Available online: https://www.avocatparis.org/actualites/responsabilite-penale-lordre-des-avocats-du-barreau-de-paris-soppose-un-projet-de-loi, (accessed on 10 July 2024).

2 See https://www.lemonde.fr/societe/article/2021/04/24/francois-molins-rien-ne-permet-d-affirmer-que-la-justice-serait-laxiste_6077883_3224.html, (accessed on 10 July 2024).

3 See https://www.conseil-etat.fr/avis-consultatifs/derniers-avis-rendus/au-gouvernement/avis-sur-un-projet-de-loi-relatif-a-la-responsabilite-penale-et-a-la-securite-interieure, (accessed on 10 July 2024).

4 See https://www.village-justice.com/articles/affaire-halimi-traore-pas-distinction-possible-selon-origine-trouble-psychique, 38890.html, (accessed on 10 July 2024).

5 See https://www.europe1.fr/politique/pas-de-proces-pour-laffaire-sarah-halimi-macron-dit-souhaiter-un-changement-de-loi-4039480, (accessed on 10 July 2024).

6 "The President of the Court of Cassation and the Prosecutor General of the Court of Cassation recall that the independence of the judiciary system, of which the President of the Republic is the guarantor, is an essential condition for the functioning of

democracy. The magistrates of the Court of Cassation must be able to examine the appeals brought before them calmly and independently." (Communiqué de la Cour de cassation, 27 January 2020).

[7] See https://www.lessurligneurs.eu/depenalisation-decriminalisation-penalisation-etc-explications, (accessed on 10 July 2024).

[8] *Club Med* is a French travel and tourism operator, specialising in the provision of all-inclusive holidays.

[9] See, https://www.lemonde.fr/societe/article/2023/01/23/face-a-la-montee-du-populisme-judiciaire-le-monde-de-la-justice-inquiet_6158965_3224.html, (accessed on 10 July 2024).

References

Allington, Daniel, Beatriz L. Buarque, and Daniel Barker Flores. 2021. Antisemitic conspiracy fantasy in the age of digital media: Three 'conspiracy theorists' and their YouTube audiences. *Language and Literature* 30: 78–102. [CrossRef]

Anesa, Patrizia. 2016. The deconstruction and reconstruction of legal information in expert-lay online interaction. *ESP Today* 4: 69–86.

Bangerter, Adrian, Pascal Wagner-Egger, and Sylvain Delouvée. 2020. How conspiracy theories spread. In *Routledge Handbook of Conspiracy Theories*. New York: Routledge, pp. 206–18.

Brès, Jacques. 2017. « Dialogisme, éléments pour l'analyse », Recherches en didactique des langues et des cultures [En ligne], 14-2|2017, mis en ligne le 15 juin 2017, consulté le 19 septembre 2024. Available online: http://journals.openedition.org/rdlc/1842 (accessed on 10 July 2024).

Boda, Zsolt, Gabriella Szabó, Attila Bartha, Gergő Medve-Bálint, and Zsuzsanna Vidra. 2015. Politically driven: Mapping political and media discourses of penal populism—The hungarian case. *East European Politics and Societies* 29: 871–91. [CrossRef]

Clément, Eloi. 2022. Loi responsabilité pénale et sécurité intérieure: Tu ne t'intoxiqueras point, Présentation du volet responsabilité pénale de la loi n° 2022-52 du 24 janvier 2022. *Dalloz actualité*. February 7. Available online: https://www.dalloz-actualite.fr/flash/loi-responsabilite-penale-et-securite-interieure-tu-ne-t-intoxiqueras-point (accessed on 10 July 2024).

Conseil d'etat, Avis consultatif. 2021. Avis sur un projet de loi relatif à la responsabilité pénale et à la sécurité intérieure Avis sur un projet de loi relatif à la responsabilité pénale et à la sécurité intérieure. July 8. Available online: https://www.conseil-etat.fr/avis-consultatifs/derniers-avis-rendus/au-gouvernement/avis-sur-un-projet-de-loi-relatif-a-la-responsabilite-penale-et-a-la-securite-interieure (accessed on 10 July 2024).

Demonceaux, Sophie. 2022. Dynamiques des échanges numériques autour d'un sujet controversé: Le cas du forum Homéopathie sur le site Doctissimo. Tic & Société [En ligne], vol. 15, N° 2–3|2ème semestre 2021—1er semestre 2022, mis en ligne le 01 juillet 2022, consulté le 20 septembre 2024. Available online: http://journals.openedition.org/ticetsociete/6865 (accessed on 10 July 2024).

Diani, Giuliana. 2022. Managing discussions in law blogs: From post to comments. *International Journal of Law, Language & Discourse* 10: 9–21.

Diani, Giuliana. 2023. Disseminating legal information on online law forums in English and Italian. *Ibérica* 46: 299–320. [CrossRef]

Douglas, Joseph E. Uscinski, Robbie M. Sutton, Aleksandra Cichocka, Turkay Nefes, Chee Siang Ang, and Farzin Deravi. 2019. Understanding conspiracy theories. *Political Psychology* 40: 3–35. [CrossRef]

Durand, Corentin. 2017. Cyberharcèlement sur le 18–25: JeuxVideo.com pourra-t-il se sauver de sa communauté. *Numerama*. Available online: https://www.numerama.com/politique/222533-cyber-harcelement-sur-le-18-25-jeuxvideo-com-pourra-t-il-se-sauver-de-sa-communaute.html (accessed on 10 July 2024).

Fovet, Thomas, Camille Lancelevée, and Pierre Thomas. 2022. Santé mentale et justice pénale en France: état des lieux et problématiques émergentes. *Bulletin de l'Académie Nationale de Médecine* 206: 301–9. [CrossRef]

Garric, Nathalie, and Julien Longhi. 2013. Atteindre l'interdiscours par la circulation des discours et du sens. *Langage et Société* 144: 65–83. [CrossRef]

Gaudin, François. 2003. *Socioterminologie: Une Approche Sociolinguistique de la Terminologie*. Bruxelles: De Boeck.

Gauducheau, Nadia, and Marcoccia Michel. 2023. La violence verbale dans un forum de discussion pour les 18–25 ans: Comment les jeunes jugent-ils les messages? *Réseaux* 241: 79–122. [CrossRef]

Hyland, Ken. 2007. Applying a gloss: Exemplifying and reformulating in academic discourse. *Applied Linguistics* 28: 266–85. [CrossRef]

Lamy, Corentin. 2017. "Jeuxvideo.com: Les coulisses du forum « 18-25 » racontées par les modérateurs". *Le Monde*. Available online: https://www.lemonde.fr/pixels/article/2017/11/16/jeuxvideo-com-les-moderateurs-racontent-les-coulisses-du-forum-18-25_5215777_4408996.html (accessed on 10 July 2024).

Longhi, Julien, and George-Elia Sarfati. 2007. Canon, doxa, vulgate: Enjeux sociodiscursifs du stéréotypage dans la dénomination intermittent. In *Stéréotypage, Stéréotypes: Fonctionnements Ordinaires et Mises en Scène*. Edited by Henri Boyer. Paris: l'Harmattan, pp. 123–31.

Makouar, Nadia. 2022. Immigration Statistics in French Online Comment Boards: Mistrust Discourse, Anti-migrant Hate Speech. In *Cyberhate in the Context of Migrations*. Cham: Springer International Publishing, pp. 115–33.

Noyon, Lucas, Jan W. De Keijser, and Jan H. Crijns. 2020. Legitimacy and public opinion: A five-step model. *International Journal of Law in Context* 16: 390–402. [CrossRef]

Plantin, Christian. 2021. Dictionnaire de l'argumentation. Available online: https://icar.cnrs.fr/dicoplantin/ (accessed on 10 July 2024).

Pratt, John. 2007. *Penal Populism*. New York: Routledge.

Philippe, Arnaud, and Aurélie Ouss. 2016. L'impact des médias sur les décisions de justice. *Notes IPP* 22: 1–5.

Rastier, François. 2011. Sémantique de corpus. In *La Mesure et le Grain*. Paris: Champion.
Rastier, François. 2017. De la sémantique structurale à la sémiotique des cultures. *Actes Sémiotiques* 120: 1–23. [CrossRef]
Rastier, François, and Nick Riemer. 2015. Interpretative semantics. In *The Routledge Handbook of Semantics*. New York: Routledge, pp. 491–506.
Salas, Denis. 2021. Justice et médias, duo ou duel? *Pouvoirs* 178: 87–96. [CrossRef]
Shahsavari, Shadi, Pavan Holur, Tianyi Wang, Timothy R. Tangherlini, and Vwani Roychowdhury. 2020. Conspiracy in the time of corona: Automatic detection of emerging COVID-19 conspiracy theories in social media and the news. *Journal of Computational Social Science* 3: 279–317. [CrossRef] [PubMed]
Turnbull, Judith Anne. 2018a. Communicating and recontextualising legal advice online in English. In *Popularization and Knowledge Mediation in the Legal Field*. Edited by Jan Engberg, Karin Luttermann, Silvia Cacchiani and Chiara Preite. Berlin: LIT Verlag, pp. 201–22.
Turnbull, Judith Anne. 2018b. "I hope somebody can help me": A linguistic analysis of British law forums. In *Frameworks for Discursive Actions and Practices of the Law*. Edited by Girolamo Tessuto, Vijay K. Bhatia and Jan Engberg. Cambridge: Cambridge Scholars, pp. 414–33.
Vicari, Stefano. 2018. Ces termes qui ne vont pas de soi ou de la circulation de la terminologie des énergies renouvelables dans les forums en ligne. *Éla. Études de Linguistique Appliquée* 192: 447–55. [CrossRef]

languages

MDPI

Article

Validation in Forensic Text Comparison: Issues and Opportunities

Shunichi Ishihara [1,*], Sonia Kulkarni [2], Michael Carne [1], Sabine Ehrhardt [3] and Andrea Nini [4]

[1] Speech and Language Laboratory, Australian National University, Canberra, ACT 2601, Australia; michael.carne@anu.edu.au
[2] School of Languages, Literatures, Cultures and Linguistics, Monash University, Clayton, VIC 3800, Australia
[3] Bundeskriminalamt, 65173 Wiesbaden, Germany; sabine.ehrhardt@bka.bund.de
[4] School of Arts, Languages and Cultures, University of Manchester, Manchester M13 9PL, UK; andrea.nini@manchester.ac.uk
* Correspondence: shunichi.ishihara@anu.edu.au

Abstract: It has been argued in forensic science that the empirical validation of a forensic inference system or methodology should be performed by replicating the conditions of the case under investigation and using data relevant to the case. This study demonstrates that the above requirement for validation is also critical in forensic text comparison (FTC); otherwise, the trier-of-fact may be misled for their final decision. Two sets of simulated experiments are performed: one fulfilling the above validation requirement and the other overlooking it, using mismatch in topics as a case study. Likelihood ratios (LRs) are calculated via a Dirichlet-multinomial model, followed by logistic-regression calibration. The derived LRs are assessed by means of the log-likelihood-ratio cost, and they are visualized using Tippett plots. Following the experimental results, this paper also attempts to describe some of the essential research required in FTC by highlighting some central issues and challenges unique to textual evidence. Any deliberations on these issues and challenges will contribute to making a scientifically defensible and demonstrably reliable FTC available.

Keywords: forensic text comparison; likelihood ratio; validation; mismatch in topics; casework conditions; relevant data

Citation: Ishihara, Shunichi, Sonia Kulkarni, Michael Carne, Sabine Ehrhardt, and Andrea Nini. 2024. Validation in Forensic Text Comparison: Issues and Opportunities. *Languages* 9: 47. https://doi.org/10.3390/languages9020047

Academic Editors: Julien Longhi and Nadia Makouar

Received: 17 September 2023
Revised: 8 January 2024
Accepted: 10 January 2024
Published: 29 January 2024

1. Introduction

1.1. Background and Aims

There is increasing agreement that a scientific approach to the analysis and interpretation of forensic evidence should consist of the following key elements (Meuwly et al. 2017; Morrison 2014, 2022):

1. The use of quantitative measurements
2. The use of statistical models
3. The use of the likelihood-ratio (LR) framework
4. Empirical validation of the method/system

These elements, it is argued, contribute towards the development of approaches that are transparent, reproducible, and intrinsically resistant to cognitive bias.

Forensic linguistic analysis (Coulthard and Johnson 2010; Coulthard et al. 2017) has been employed for analyzing documents as forensic evidence[1] to infer the source of a questioned document (Grant 2007, 2010; McMenamin 2001, 2002). Indeed, this has been crucial in solving several cases; see e.g., Coulthard et al. (2017). However, analyses based on an expert linguist's opinion have been criticized for lacking validation (Juola 2021). Even where textual evidence is measured quantitatively and analyzed statistically, the interpretation of the analysis has rarely been based on the LR framework (c.f., Ishihara 2017, 2021, 2023; Ishihara and Carne 2022; Nini 2023).

The lack of validation has been a serious drawback of forensic linguistic approaches to authorship attribution. However, there is a growing acknowledgment of the importance of

validation in this field (Ainsworth and Juola 2019; Grant 2022; Juola 2021); This acknowledgment is fully endorsed. That being said, to the best of our knowledge, the community has not started thinking in depth as to what empirical validation obliges us to do. Looking at other areas of forensic science, there is already some degree of consensus on how empirical variation should be implemented (Forensic Science Regulator 2021; Morrison 2022; Morrison et al. 2021; President's Council of Advisors on Science and Technology (U.S.) 2016). In forensic science more broadly, two main requirements[2] for empirical validation are:

- Requirement 1: reflecting the conditions of the case under investigation;
- Requirement 2: using relevant data to the case.

The current study stresses that these requirements are also important in the analysis of forensic authorship evidence. This is demonstrated by comparing the results of the two competing types of experiments, one satisfying the above requirements and the other disregarding them.

The LR framework is employed in this study. LRs are calculated using a statistical model from the quantitatively measured properties of documents.

Real forensic texts have a mismatch or mismatches in topics, so this is the casework condition for which we will select relevant data. Amongst other factors, mismatch in topics is typically considered a challenging factor in authorship analysis (Kestemont et al. 2020; Kestemont et al. 2018). Cross-topic or cross-domain comparison is an adverse condition often used in the authorship attribution/verification challenges organized by PAN.[3]

Following the experimental results, this paper also describes future research necessary for forensic text comparison (FTC)[4] by highlighting some crucial issues and challenges unique to the validation of textual evidence. These include (1) determining specific casework conditions and mismatch types that require validation; (2) determining what constitutes relevant data; and (3) the quality and quantity of data required for validation.

1.2. Likelihood-Ratio Framework

The LR framework has long been argued to be the logically and legally correct approach for evaluating forensic evidence (Aitken and Taroni 2004; Good 1991; Robertson et al. 2016) and it has received growing support from the relevant scientific and professional associations (Aitken et al. 2010; Association of Forensic Science Providers 2009; Ballantyne et al. 2017; Forensic Science Regulator 2021; Kafadar et al. 2019; Willis et al. 2015). In the United Kingdom, for instance, the LR framework will need to be deployed in all of the main forensic science disciplines by October 2026 (Forensic Science Regulator 2021).

An LR is a quantitative statement of the strength of evidence (Aitken et al. 2010), as expressed in Equation (1).

$$\mathrm{LR} = \frac{p(E|H_p)}{p(E|H_d)} \tag{1}$$

In Equation (1), the LR is equal to the probability (p) of the given evidence (E) assuming that the prosecution hypothesis (H_p) is true, divided by the probability of the same evidence assuming that the defense hypothesis (H_d) is true. The two probabilities can also be interpreted, respectively, as *similarity* (how similar the samples are) and *typicality* (how distinctive this similarity is). In the context of FTC, the typical H_p is that "the source-questioned and source-known documents were produced by the same author" or "the defendant produced the source-questioned document". The typical H_d is that "the source-questioned and source-known documents were produced by different individuals" or "the defendant did not produce the source-questioned document".

If the two probabilities are the same, then the LR = 1. If, however, $p(E|H_p)$ is larger than $p(E|H_d)$, then the LR will be larger than one and this means that there is support for H_p. If, instead, $p(E|H_d)$ is larger than $p(E|H_p)$, then an LR < 1 will indicate that there is more support for H_d (Evett et al. 2000; Robertson et al. 2016). The further away from one, the more strongly the LR supports either of the competing hypotheses. An LR of ten, for

example, should be interpreted as the evidence being ten times more likely to be observed assuming the H_p being true than assuming the H_d being true.

The belief of the trier-of-fact regarding the hypotheses was possibly formed by previously presented evidence, and logically it should be updated by the LR. In a layperson's term, that is, the belief of the decision maker regarding the suspect being guilty or not changes as a new piece of evidence is presented to them. This process is formally expressed in Equation (2).

$$\underbrace{\frac{p(H_p)}{p(H_d)}}_{prior\ odds} \times \underbrace{\frac{p(E|H_p)}{p(E|H_d)}}_{LR} = \underbrace{\frac{p(H_p|E)}{p(H_d|E)}}_{posterior\ odds} \tag{2}$$

Equation (2) is the so-called odds form of Bayes' Theorem. It states that the multiplication of the *prior odds* and the *LR* equates to the *posterior odds*. The prior odds is the belief of the trier-of-fact with respect to the probability of the H_p or H_d being true, before the LR of a new piece of evidence is presented. The posterior odds quantifies the up-to-date belief of the trier-of-fact after the LR of the new evidence is presented.

As Equation (2) shows, calculation of the posterior odds requires both the prior odds and the LR. Thus, it is logically impossible for a forensic scientist to compute the posterior odds during their evidential analysis because they are not in the position of knowing the trier-of-fact's belief. It is legally inappropriate for the forensic practitioner to present the posterior odds because the posterior odds concerns the ultimate issue of the suspect being guilty or not (Lynch and McNally 2003). If they do so, the forensic scientist deviates from their authority.

1.3. Complexity of Textual Evidence

Besides linguistic-communicative contents, various other pieces of information are encoded in texts. These may include information about (1) the authorship; (2) the social group or community the author belongs to; (3) the communicative situations under which the text was composed, and so on (McMenamin 2002). Every author or individual has their own 'idiolect': a distinctive individuating way of speaking and writing (McMenamin 2002). This concept of idiolect is fully compatible with modern theories of language processing in cognitive psychology and cognitive linguistics, as explained in Nini (2023).

The 'group-level' information that is associated with texts can be collated for the purpose of author profiling (Koppel et al. 2002; López-Monroy et al. 2015). The group-level information may include: the gender, age, ethnicity, and social-economical background of the author.

The writing style of each individual may vary depending on communicative situations that may be a function of internal and external factors. Some examples are the genre, topic, and level of formality of the texts; the emotional state of the author; and the recipient of the text.

As a result, a text is a reflection of the complex nature of human activities. As introduced in Section 1.1, we only focus on topic as a source of mismatch. However, topic is only one of many potential factors that influence individuals' writing styles. Thus, in real casework, the mismatch between the documents under comparison is highly variable; consequently, it is highly case specific. This point is further discussed in Section 7.

2. Database and Setting up Mismatches in Topics

Taking up the problem of mismatched topics between the source-questioned and source-known documents as a case study, this study demonstrates that validation experiments should be performed by (1) reflecting the conditions of the case under investigation and (2) using data relevant to the case.

2.1. Database

The Amazon Authorship Verification Corpus (AAVC) http://bit.ly/1OjFRhJ (accessed on 30 September 2020) (Halvani et al. 2017) was used in this study. The AAVC contains reviews on Amazon products submitted by 3227 authors. As can be seen from Figure 1 which shows the number of reviews contributed by the authors, five or more reviews were collected from the majority of reviewers included in the AAVC. Altogether 21,347 reviews are included in the AAVC.

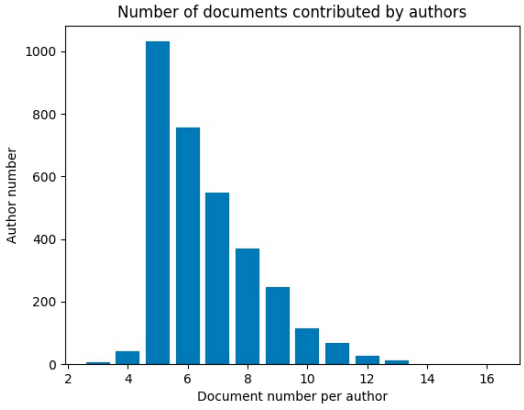


Figure 1. Number of reviews (documents) contributed by authors.

The reviews are classified into 17 different categories as presented in Figure 2. In the AAVC, each review is equalized to 4 kB, which is approximately 700–800 words in length.

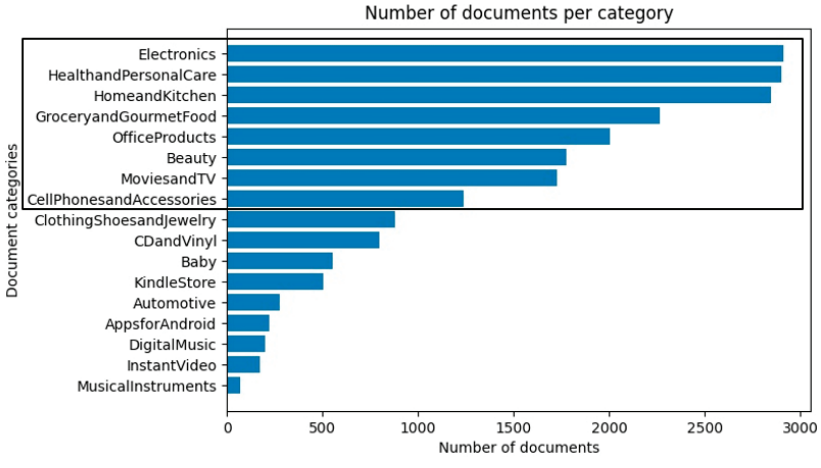


Figure 2. The 17 review categories of the AAVC and their numbers of reviews. The categories that have the most reviews (top eight) are indicated by a black rectangle.

These reviews and the categories of the AAVC are referred to from now on as "documents" and "topics", respectively.

The AAVC is a widely recognized corpus specifically designed for authorship verification studies, as evidenced by its utilization in various studies (Boenninghoff et al. 2019; Halvani et al. 2020; Ishihara 2023; Rivera-Soto et al. 2021). Certain aspects of the data, such as genre and document length, are well-controlled. However, there are uncontrolled variables that may bear relevance to the outcomes of the current study. For instance, there

is no control over the input device used by reviewers (e.g., mobile device or computer) (Murthy et al. 2015), the English variety employed, or whether writing assistance functions such as automatic spelling and grammar checkers have been activated. All of these factors are likely to influence the writing style of individuals. Furthermore, the corpus may include some fake reviews, as the same user ID might be used by multiple reviewers, and conversely, the same reviewer may use multiple user IDs. Nonetheless, considering realistic forensic conditions, it is practically impossible to exert complete control over the data. Multiple corpora are often employed in authorship studies, investigating the robustness of systems across a variety of data. To the best of our knowledge, no peculiar behavior of the AAVC has been reported in any studies, ascertaining the quality of the corpus to the appropriate extent.

The topic categories employed in the AAVC appear to be somewhat arbitrary, with certain topics seemingly not situated at the same hierarchical level; for instance, "Cell Phones and Accessories" could be considered a subcategory of "Electronics". Partially owing to overlaps across some topics, Section 2.2 will illustrate that documents belonging to certain topics exhibit similar patterns of distribution. Nevertheless, Section 2.2 also reveals that documents in some topics showcase unique distributional patterns distinct from other topics, and these topics are utilized for simulating topic mismatches.

2.2. Distributional Patterns of Documents Belonging to Different Topics

In order to show the similarities (or differences) between documents and topics, documents belonging to the top eight most frequent topics, which are indicated by a rectangle in Figure 2, are plotted in a two-dimensional space using t-distributed stochastic neighbor embedding (T-SNE)[5] (van der Maaten and Hinton 2008) in Figure 3. Prior to the T-SNE, each document was vectorized via a transformer-based large language model, BERT[6] (Devlin et al. 2019). Vectorization or word embedding is the process of converting texts to numerical vectors, which are high in dimension. In this way, each document is holistically represented in a semantically deep manner. Yet, it is difficult to visualize the high-dimensional data. T-SNE allows the visualization of high-dimensional data by reducing the dimension in a non-linear manner. T-SNE is a commonly used dimension reduction technique in which the text data are represented with word embeddings. This is because T-SNE is known to preserve the local and global relationships of data even after dimension reduction (van der Maaten and Hinton 2008). Thus, Figure 3 is considered to effectively depict the actual differences and similarities between the documents included in the different topics.

In Figure 3, each point represents a separate document. The distances between the points reflect the degrees of similarity or difference between the corresponding documents. Some topics have more points than others, reflecting the different numbers of documents included in the topics[7] (see Figure 2). A red-filled circle in each plot indicates the centroid (the mean T-SNE values of Dimensions 1 and 2) of the documents belonging to the topic.

The documents belonging to the eight different topics display some unique distributional patterns; e.g., some topics show a similar distributional pattern to each other while other topics display their own unique patterns. The documents categorized into "Office Products", "Electronics", "Home and Kitchen", and "Health and Personal Care" are similar to each other in that they are most widely distributed in the space; consequently, they extensively overlap each other. That is, there are a wide variety of documents included in these topics. The similarity of these four topics can also be seen from the fact that the centroids are all located in the middle of the plots. The documents in the "Beauty", "Grocery and Gourmet Food", "Movies and TV", and "Cellphones and Accessories" topics are more locally distributed and their areas of concentrations are rather different. In particular, the documents in the "Beauty" and "Movie and TV" topics are most clustered in different areas; as a result, the centroids appear in different locations. That is, those documents belonging to each of the "Beauty" and "Movie and TV" topics are less diverse within each topic, but they are largely different from each other.

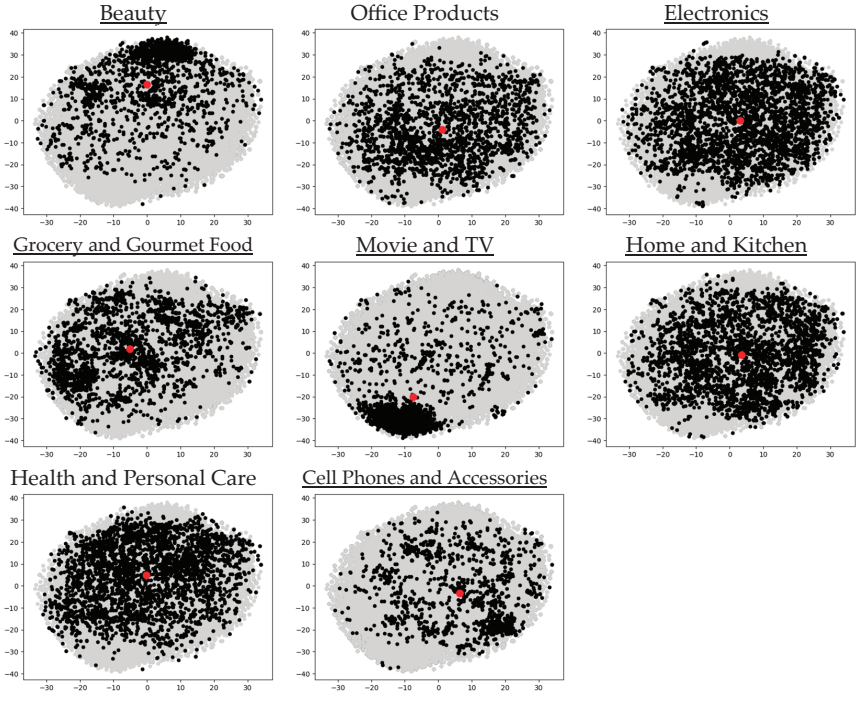


Figure 3. T-SNE plots of the documents belonging to the eight topics indicated in Figure 2. The underlined topics are used for simulating the mismatches in topics. The red-filled circle in each plot shows the centroid of the documents belonging to the topic. x-axis = Dimension 1; y-axis = Dimension 2.

Primarily focusing on the overall distances between documents belonging to different topics, mismatches in topics were simulated in Section 2.3, varying in the degree of distance. Specifically, in these simulated mismatches, the degree of distance between the two centroids differs. Figure 3 illustrates that, in addition to the centroids' locations, documents classified under different topics display diverse distributional patterns, with some being more dispersed or clustered than others. These distinctive distributional patterns may influence the experimental results, including LR values. However, the consideration of these patterns was limited primarily due to the difficulties associated with simulation.

2.3. Simulating Mismatch in Topics

Judging from the distributional patterns that can be observed from Figure 3 for the eight topics, the following three cross-topic settings were used for the experiments, together with paired documents that were randomly selected without considering their topic categories (Any-topics).

- Cross-topic 1: "Beauty" vs. "Movie and TV"
- Cross-topic 2: "Grocery and Gourmet Food" vs. "Cell Phones and Accessories"
- Cross-topic 3: "Home and Kitchen" vs. "Electronics"
- Any-topics: Any-topic vs. Any-topic

Cross-topics 1, 2, and 3 display different degrees of dissimilarity between the paired topics, which are visually observable in Figure 4.

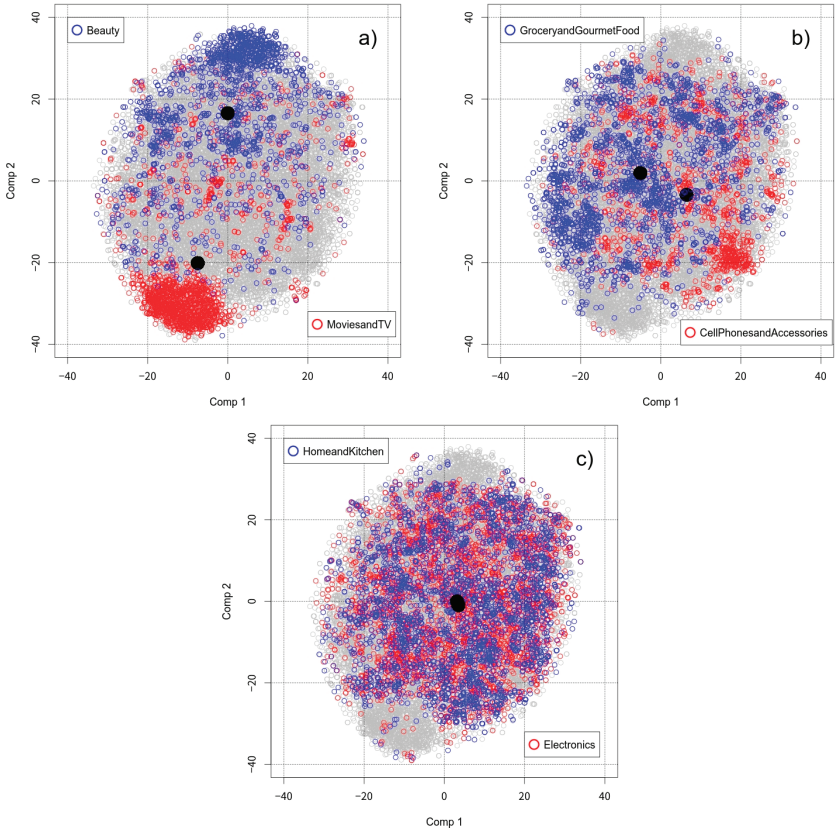


Figure 4. Combined T-SNE plots for Cross-topics 1 (Panel **a**), 2 (Panel **b**), and 3 (Panel **c**), respectively. Black-filled circles in each panel show the centroids of the paired topics.

The documents classified as "Beauty" or "Movie and TV" (Cross-topic 1) show the greatest distances between the documents of the topics (see Figure 4a) in their distributions. The centroids of the documents for each topic, indicated by the black points, are far apart in Figure 4a. It can be foreseen that this large gap observed in Cross-topic 1 will make the FTC challenging. On the other hand, the documents classified as "Home and Kitchen" and "Electronics" (Cross-topic 3) heavily overlap each other in their distributions (see Figure 4c); the centroids are very closely located to each other, so it is likely that this FTC will be less challenging than for Cross-topic 1. Cross-topic 2 is somewhat in-between Cross-topic 1 and Cross-topic 3 in terms of the degree of overlap between the documents belonging to the "Grocery and Gourmet Food" and "Cell Phones and Accessories" topics.

The documents belonging to the "Any-topics" category were randomly selected from the AAVC.

Altogether, 1776 same-author (SA) and 1776 different-author (DA) pairs of documents were generated for each of the four settings given in the bullets above, and they were further partitioned into six mutually exclusive batches for cross-validation experiments. That is, 296 (=1776 ÷ 6) SA and 296 (=1776 ÷ 6) DA unique comparisons are included in each batch of the four settings. Refer to Section 3.1 for detailed information on data partitioning and the utilization of these batches in the experiments.

As can be seen from Figure 2, the number of documents included in each of the selected six topics is different; thus, the maximum numbers of paired documents for SA comparisons are also different between the three Cross-topics. The number of possible SA

comparisons is 1776 for Cross-topic 2, and this is the smallest out of the three Cross-topics. Thus, the number of the SA comparisons is equalized to 1776 also for Cross-topics 1 and 3 by a random selection. The number of DA comparisons is also matched with that of SA comparisons. 1776 DA comparisons were randomly selected from all possible DA comparisons in such a way that each of the 1776 DA comparisons has a unique combination of authors.

Focusing on the mismatch in topics, two simulated experiments (Experiments 1 and 2) were prepared with the described subsets of the AAVC. Experiments 1 and 2 focus on Requirements 1 and 2, respectively. Experiments 1 and 2 each further include two types of experiments: one fulfilling the requirement and the other overlooking it. Detailed structures of the experiments will be described in Section 4.

3. Calculating Likelihood Ratios: Pipeline

After representing each document as a vector comprising a set of features, calculating an LR for a pair of documents under comparison (e.g., source-questioned and source-known documents) is a two-stage process consisting of the score-calculation stage and the calibration stage. The pipeline for calculating LRs is shown in Figure 5 for validation of the FTC system.

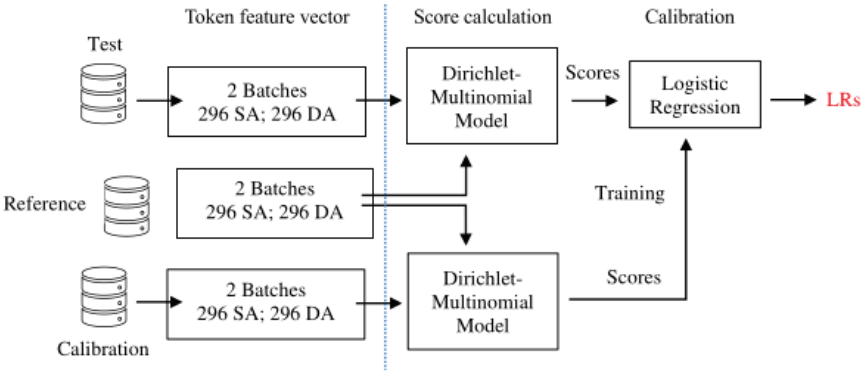


Figure 5. Schematic illustration of the process for likelihood ratio calculations.

Details of the partitioned databases and the stages of the pipeline are provided in the following sub-sections.

3.1. Database Partitioning

As can be seen from Figure 5, three mutually exclusive databases are necessary for validating the performance of the LR-based FTC system. They are the Test, Reference, and Calibration databases. Using two independent batches (out of six) at a time for each of the Test, Reference, and Calibration databases, six cross-validation experiments are possible as shown in Table 1.

Table 1. Use of the batches for the Test, Reference, and Calibration databases.

Experiments	Test	Reference	Calibration
Experiment 1	Batch 1, Batch 2	Batch 3, Batch 4	Batch 5, Batch 6
Experiment 2	Batch 2, Batch 3	Batch 4, Batch 5	Batch 6, Batch 1
Experiment 3	Batch 3, Batch 4	Batch 5, Batch 6	Batch 1, Batch 2
Experiment 4	Batch 4, Batch 5	Batch 6, Batch 1	Batch 2, Batch 3
Experiment 5	Batch 5, Batch 6	Batch 1, Batch 2	Batch 3, Batch 4
Experiment 6	Batch 6, Batch 1	Batch 2, Batch 3	Batch 4, Batch 5

The SA and DA comparisons included in the Test database are used for assessing the performance of the FTC system. In the first stage of the pipeline given in Figure 5 (the score-calculation stage), a score is estimated for each comparison generated from the Test database, considering the similarity between the documents under comparison as well as the typicality of them. For assessing typicality, the necessary statistical information was obtained from the Reference database.

As two batches are used for each database in each of the six cross-validation experiments, 592 (=296 × 2) SA scores and 592 (=296 × 2) DA scores are obtained for each experiment. These 592 SA scores and 592 DA scores from the Test database are converted to LRs at the following stage of calibration. Scores are also calculated for the SA and DA comparisons from the Calibration database; that is, 592 SA and 592 DA scores for each experiment. These scores from the Calibration database are used to convert the scores of the Test database to LRs. For the explication of score calculation and calibration, see Sections 3.3 and 3.4, respectively.

3.2. Tokenization and Representation

Each document was word-tokenized using the token() function of the quanteda R library (Benoit et al. 2018); the default settings were applied. Note that this tokenizer recognizes punctuation marks (e.g., '?', '!', and '.') and special characters (e.g., '$', '&', and '%') as independent words; thus, they constitute tokens by themselves. No stemming algorithm is applied. Upper and lower cases are treated separately; that is, 'book' and 'Book' are treated as separate words. It is known that the use of upper/lower case characters is fairly idiosyncratic (Zhang et al. 2014).

Each document of the AAVC is bag-of-words modelled with the 140 most frequent tokens appearing in the entire corpus, which are listed in Table A1 of Appendix A. The reader can verify that these are common words, being used regardless of topics. Obvious topic-specific words start appearing if the list of words is further extended.

An example the bag-of-words model is given in Example 1.

Example 1. $Document = \begin{Bmatrix} T_1 & T_2 & T_3 & \cdots, & T_{139} & T_{140} \\ 21' & 19' & 13' & \cdots, & 0' & 1 \end{Bmatrix}$.

The top 15 tokens are shown in Table 2 along with their occurrences. That is, these 15 tokens constitute the first 15 items of the bag-of-words feature vector: from T_1 to T_{15} for Example 1. As could be expected, many of the tokens included in Table 2 are function words and punctuation marks.

Table 2. Occurrences of the 15 most frequent tokens in the entire AAVC.

Rank	Token	Occurrences
1	.	829,646
2	the	678,218
3	“ (open)	651,324
4	” (close)	439,226
5	I	426,411
6	and	424,477
7	a	401,828
8	to	275,859
9	it	275,463
10	of	267,951
11	is	182,605
12	for	174,646
13	that	173,641
14	in	170,556
15	this	133,056

Many stylometric features have been developed to quantify writing style. Stamatatos (2009) classifies stylometric features into the five different categories of 'lexical', 'character', 'syntactic', 'semantic', and 'application-specific' and summarizes their pros and cons. It may be that different features have different degrees of tolerance to different types of mismatches. Thus, different features should be selectively used according to the casework conditions. However, it is not an easy task to unravel the relationships between them. Partly because of this, it is a common practice for the cross-domain authorship verification systems built on traditional feature engineering to use an ensemble of different feature types (Kestemont et al. 2020; Kestemont et al. 2021).

3.3. Score Calculation

The bag-of-words model consists of token counts, so the measured values are discrete. As such, the Dirichlet-multinomial statistical model was used to calculate scores. The effectiveness and appropriateness of the model for authorship-textual evidence has been demonstrated in Ishihara (2023). The formula for calculating a score for the source-questioned (X) and source-known (Y) documents with the Dirichlet-multinomial model is given in Equation (A1) of Appendix A. In essence, taking into account the discrete nature of the measured feature values, the model evaluates the similarity between X and Y and their typicality against the samples included in the Reference database, calculating a score. With the level of typicality held constant, the more similar X and Y are, the higher the score will be. Conversely, for an identical level of similarity, the more typical X and Y are, the smaller the score will become.

3.4. Calibration

The score obtained at the score-calculation stage for a pair of documents is LR-like in that it reveals the degree of similarity between the documents while considering the typicality with respect to the relevant population. However, if the Dirichlet-multinomial statistical model does not return well-calibrated outputs, they cannot be interpreted as LRs. In fact, this is often the case.[8] This point is illustrated in Figure 6 using the distributions of imaginary SA and DA scores/LRs.

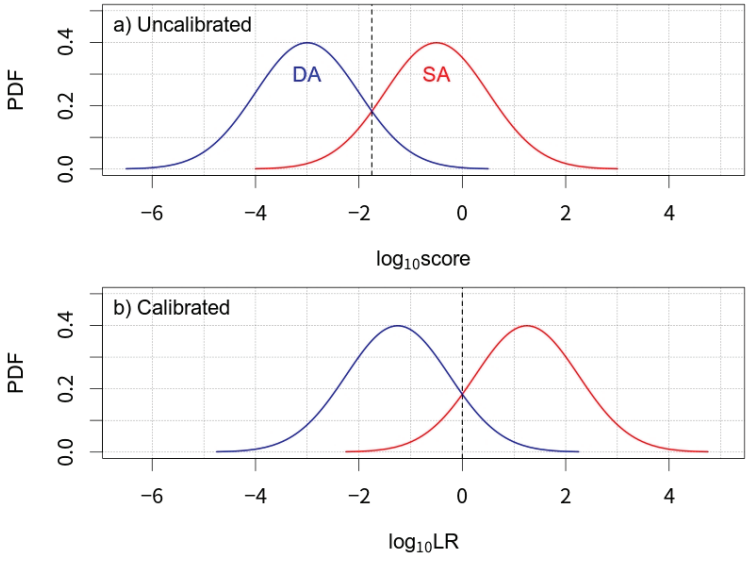


Figure 6. Schematic illustration of the concept of calibration. SA and DA are the example outputs of a system for same-author and different-author comparisons, respectively; (**a**,**b**) are uncalibrated and calibrated systems, respectively. PDF—Probability density function.

In Figure 6a, the neutral point that optimally separates the DA and SA comparisons (the vertical dashed line of Figure 6a), is not aligned with a $\log_{10}$ value of 0, which is the neutral point in the LR framework. In such a case, the calculated value cannot be translated as the strength of the evidence. Thus, it is customarily called a 'score' (uncalibrated LR). Figure 6b is an example case of a calibrated system.

The scores (uncalibrated LRs) need to be calibrated or converted to LRs. Logistic regression is the standard method for this conversion (Morrison 2013). In other words, the scores of the Calibration database are used to train logistic regression for calibration.

Calibration is integral to the LR framework, as raw scores can be misleading until converted to LRs. Readers are encouraged to explore (Morrison 2013, 2018; Ramos and Gonzalez-Rodriguez 2013; Ramos et al. 2021) for a deeper understanding of the significance of calibration in evaluating evidential strength.

4. Experimental Design: Reflecting Casework Conditions and Using Relevant Data

Regarding the two requirements (Requirements 1 and 2) for validation stated in Section 1.1, two experiments (Experiments 1 and 2) were designed under cross-topic conditions. In the experiments, Cross-topic 1 is assumed to be the casework condition in which the source-questioned text is written on "Beauty" and the source-known text is written on "Movie and TV". Readers will recall (Section 2.2) that Cross-topic 1 has a high degree of topic mismatch. In order to conduct validation under the casework conditions with the relevant data, the validation experiment should be performed with the databases having pairs of documents reflecting the same mismatch in topics as Cross-topic 1. Figure 7 elucidates this.

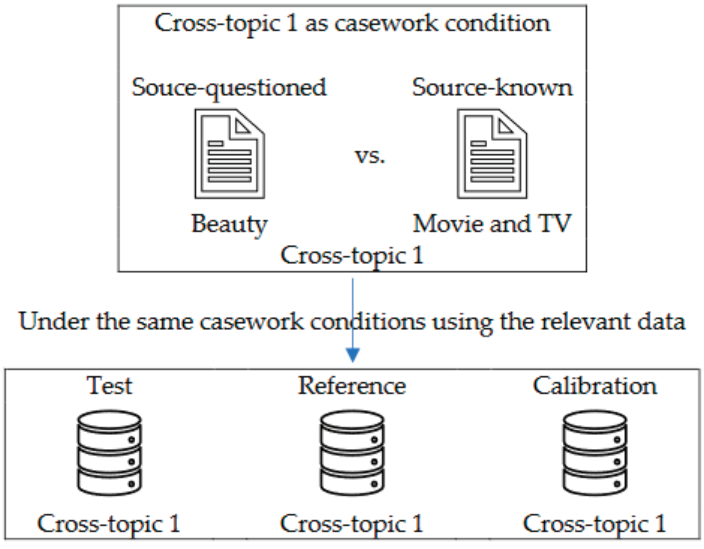


Figure 7. Example illustrating validation under the same conditions as the casework with the relevant data.

Sections 4.1 and 4.2 explain how Experiments 1 and 2 were set up, respectively. Experiment 1 considers Requirement 1 for validation, and Experiment 2 considers Requirement 2 for relevant data.

4.1. Experiment 1: Fulfilling or Not Fulfilling Casework Conditions

If the casework condition illustrated in Figure 7 were to be ignored, the validation experiment would be performed using Cross-topic 2, Cross-topic 3, or Any-topic. This is summarized in Table 3.

Table 3. Conditions of Experiment 1.

	Test	Reference	Calibration
Under casework condition	Cross-topic 1	Cross-topic 1	Cross-topic 1
Not under casework condition	Cross-topic 2	Cross-topic 2	Cross-topic 2
Not under casework condition	Cross-topic 3	Cross-topic 3	Cross-topic 3
Not under casework condition	Any-topic	Any-topic	Any-topic

The results of the validation experiments carried out under the conditions specified in Table 3 are presented and compared in Section 6.1.

4.2. Experiment 2: Using or Not Using Relevant Data

If data relevant to the case were not used for calculating the LR for the source-questioned and source-known documents under investigation, what would happen to the LR value? This question is the basis of Experiment 2. As such, in Experiment 2, validation experiments were carried out with the Reference and Calibration databases, which do not share the same type of topic mismatch as the Test database (Cross-topic 1). Table 4 includes the conditions used in Experiment 2.

Table 4. Conditions of Experiment 2.

	Test	Reference	Calibration
Using the relevant data	Cross-topic 1	Cross-topic 1	Cross-topic 1
Not using the relevant data	Cross-topic 1	Cross-topic 2	Cross-topic 2
Not using the relevant data	Cross-topic 1	Cross-topic 3	Cross-topic 3
Not using the relevant data	Cross-topic 1	Any-topic	Any-topic

The results of the validation experiments carried out under the conditions specified in Table 4 are presented and compared in Section 6.2.

5. Assessment

The performance of a source-identification system is commonly assessed in terms of its identification accuracy and/or identification error rate. Metrics such as precision, recall, and equal error rate are typical in this context. However, these metrics are not appropriate for evaluating LR-based inference systems (Morrison 2011, p. 93). These metrics are based on the binary decision of whether the identification is correct or not, which is implicitly tied to the ultimate issue of the suspect being deemed guilty or not guilty. As explained in Section 1.2, forensic scientists should refrain from making references to this matter. Furthermore, these metrics fail to capture the gradient nature of LRs; they do not take into account the actual strength inherent in these ratios.

The performance of the FTC system was assessed by means of the log-likelihood-ratio cost (C_{llr}), which was first proposed by Brümmer and du Preez (2006). It serves as the conventional assessment metric for LR-based inference systems. The C_{llr} is described in detail in van Leeuwen and Brümmer (2007) and Ramos and Gonzalez-Rodriguez (2013). Equation (A2) of Appendix A is for calculating C_{llr}. An example of the C_{llr} calculation is also provided in Appendix A.

In the calculation of C_{llr}, each LR value attracts a certain cost.[9] In general, the contrary-to-fact LRs; i.e., LR < 1 for SA comparisons and LR > 1 for DA comparisons, are assigned far more substantial costs than the consistent-with-fact LRs; i.e., LR > 1 for SA comparisons and LR < 1 for DA comparisons. For contrary-to-fact LRs, the cost increases as they are farther away from unity. For consistent-with-fact LRs, the cost increases as they become closer to unity. The C_{llr} is the overall average of the costs calculated for all LRs of a given experiment. See Appendix C.2 of Morrison et al. (2021) for the different cost functions of the consistent-with-fact and contrary-to-fact LRs.

The C_{llr} is a metric assessing the overall performance of an LR-based system. The C_{llr} consists of two metrics that assess the discrimination performance and calibration performance of the system, respectively. They are called discrimination loss (C_{llr}^{min}) and calibration loss (C_{llr}^{cal}). The C_{llr}^{min} is obtained by calculating the C_{llr} for the optimized LRs via the non-parametric pool-adjacent-violators algorithm. The difference between C_{llr}^{min} and C_{llr} is C_{llr}^{cal}; i.e., $C_{llr} = C_{llr}^{min} + C_{llr}^{cal}$. If we consider the cases presented in Figure 6 as examples for C_{llr}, C_{llr}^{min}, and C_{llr}^{cal}, the discriminating potential of the system in Figure 6a and that in Figure 6b are the same. In other words, the C_{llr}^{min} values for both are identical. The distinction lies in their C_{llr}^{cal} values, where the C_{llr}^{cal} value of Figure 6a should be higher than that of Figure 6b. Consequently, the overall C_{llr} will be higher for Figure 6a than for Figure 6b.

More detailed descriptions of these metrics can be found in Brümmer and du Preez (2006), Drygajlo et al. (2015) and in Meuwly et al. (2017).

A C_{llr} less than one means that the system provides useful information for discriminating between authors. The lower the C_{llr} value, therefore, the better the system performance. This holds true for both C_{llr}^{min} and C_{llr}, concerning the system's discrimination and calibration performances.

The derived LRs are visualized by means of Tippett plots. A description of Tippett plots is given in Section 6.2., in which the LRs of some experiments are presented.

6. Results

The results of Experiments 1 and 2 are separately presented in Sections 6.1 and 6.2. The reader is reminded that in each experiment, six cross-validated experiments were performed separately for each of the four conditions specified in Table 3 (Experiment 1) and Table 4 (Experiment 2), and also that Cross-topic 1, which has a large topic mismatch, is presumed to be the casework condition in which the source-questioned text is written on "Beauty" and the source-known text is written on "Movie and TV".

6.1. Experiment 1

In Figure 8, the maximum, mean and minimum C_{llr} values of the six experiments are plotted for the four conditions given in Table 3. Please recall that the lower in C_{llr}, the better the performance.

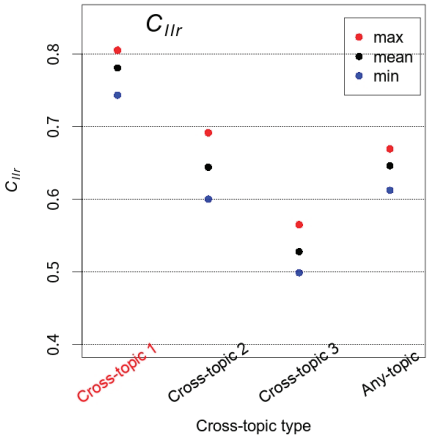


Figure 8. The maximum (max), mean, and minimum (min) C_{llr} values of the six cross-validated experiments are plotted for each of the four experimental conditions specified in Table 3.

Regarding the degree of mismatch in topics that was described in Section 2.3, the experiment with Cross-topic 1, which matches the casework condition, yielded the worst

performance result (mean C_{llr} = 0.78085) while the experiment with Cross-topic 3 yielded the best (mean C_{llr} = 0.52785). The experiments with Cross-topic 2 (mean C_{llr} = 0.65643) and Any-topic (mean C_{llr} = 0.64412) came somewhere in-between Cross-topics 1 and 3. It appears that the FTC system provides some useful information regardless of the experimental conditions; the C_{llr} values are all smaller than one. However, fact-finders would be led to believe that the performance of the FTC system is better than it actually is if they were informed with the validation result that does not match the casework condition; namely, Cross-topics 2 and 3 and Any-topic. Obviously, the opposite instance is equally likely in which an FTC system is judged to be worse than it actually is.

One may think it sensible to validate the system under less-constrained or more-inclusive heterogeneous conditions. However, the experimental result with Any-topic demonstrated that this was not appropriate, since the FTC system clearly performed differently from the experiment that was conducted under the same condition as the casework condition.

The performance of the FTC system is further analyzed by looking into its discrimination and calibration costs independently. The C_{llr}^{min} and C_{llr}^{cal} are plotted one by one in Panels (a) and (b) of Figure 9 for the same experimental conditions listed in Table 3.

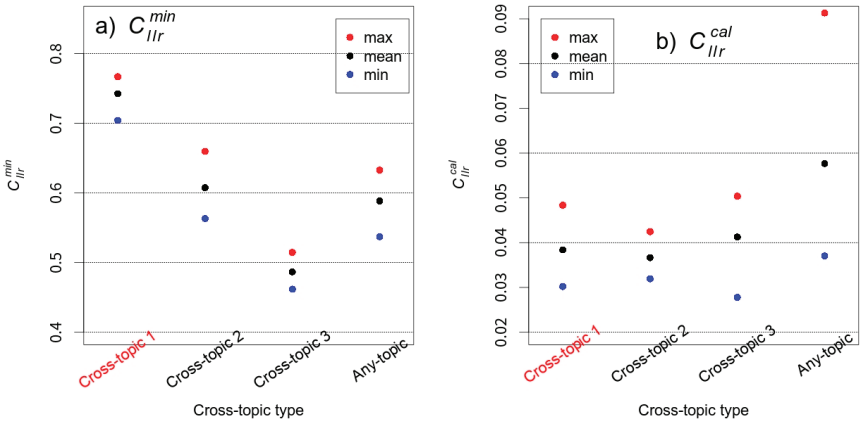


Figure 9. The maximum (max), mean, and minimum (min) C_{llr}^{min} (Panel **a**) and C_{llr}^{cal} (Panel **b**) values of the six cross-validated experiments are plotted for the four experimental conditions specified in Table 3.

The differences in discrimination performance (measured in C_{llr}^{min}) observed in Figure 9a between the four conditions is parallel to the differences in overall performance (measured in C_{llr}) observed in Figure 8 between the same four conditions. That is, the discrimination between the SA and DA documents is more challenging for one cross-topic type than another. The difficulty is in the descending order of Cross-topic 1, Cross-topic 2, and Cross-topic 3. The discrimination performance of Any-topic is marginally better than that of Cross-topic 2.

The C_{llr}^{cal} values charted in Figure 9b are all close to zero and they are similar to each other; note that the range of the y-axis is very narrow between 0.02 and 0.09. That is to say, the resultant LRs are all well-calibrated. However, it appears that Any-topic (mean C_{llr}^{cal} = 0.05766) underperforms the other Cross-topic types in calibration performance. The calibration performances of Cross-topics 1, 2 and 3 are virtually the same (mean C_{llr}^{cal} is 0.03839 for Cross-topic 1; 0.03664 for Cross-topic 2; and 0.04126 for Cross-topic 3). As explained in Section 2.3, paired documents belonging to Any-topic were randomly selected from the entire database, which allows large variability between the batches. This could be a possible reason for the marginally larger C_{llr}^{cal} values for Any-topic. However, this warrants further investigation.

6.2. Experiment 2

The maximum, mean and minimum C_{llr} values of the six experiments are plotted separately in Figure 10 for each of the four conditions specified in Table 4.

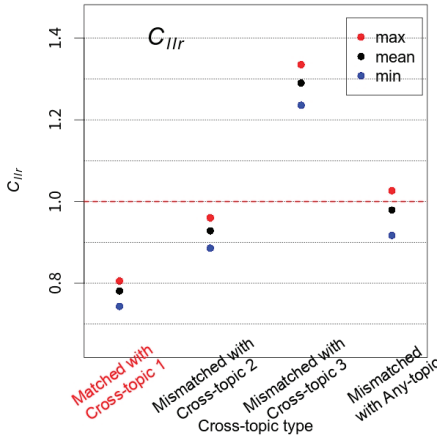


Figure 10. The maximum (max), mean, and minimum (min) C_{llr} values of the six cross-validated experiments are plotted for each of the four experimental conditions specified in Table 4. The red horizontal-dashed line indicates $C_{llr} = 1$.

The experimental results given in Figure 10 clearly show that it is detrimental to calculate LRs with data that is irrelevant to the case. The C_{llr} values can go beyond one, i.e., the system is not providing useful information for the case. The degree of deterioration in performance depends on the Cross-topic types used for Reference and Calibration databases. Cross-topic 3, which has the greatest difference from Cross-topic 1 (compare Figures 4a and 4c), caused more substantial impediment in performance in comparison to Cross-topic 2. It is also interesting to see that the use of Any-topic for Reference and Calibration databases, which may be considered the most generic dataset reflecting the overall characteristics of the entire database, also brought about a decline in performance, as the C_{llr} values can go over one. The results included in Figure 10 well demonstrate the risk of using irrelevant data, i.e., the degree of topic mismatch is not comparable between Test and Reference/Calibration databases for calculating LRs. This may result in jeopardizing the genuine value of the evidence.

In order to further investigate the cause of the deterioration in overall performance (measured in C_{llr}), the C_{llr}^{min} and C_{llr}^{cal} are plotted in Figure 11 in the same manner as in Figure 9. Panels (a) and (b) are for the C_{llr}^{min} and C_{llr}^{cal}, respectively.

Panel (a) of Figure 11 shows that the discrimination performance evaluated by C_{llr}^{min} is effectively the same across all of the experimental conditions (C_{llr}^{min} mean: 0.74246 for Cross-topic 1; 0.73786 for Cross-topic 2; 0.73618 for Cross-topic 3; and 0.74102 for Any-topic). That is, as far as the discriminating power is concerned, the degree of mismatch in topics does not result in any sizable difference in discriminability. The discriminating power of the system remains unchanged before and after calibration, i.e., the C_{llr}^{min} value does not change before and after calibration. Since the Calibration database does not cause variability, and the Test database is fixed to Cross-topic 1, only the Reference database plays a role in the variability of the C_{llr}^{min} values across the four experimental conditions. The results given in Figure 11a imply that not using the relevant data for the Reference database does not have an apparently negative impact on the discriminability of the system. This point will be discussed further after the results of C_{llr}^{cal} are presented below.

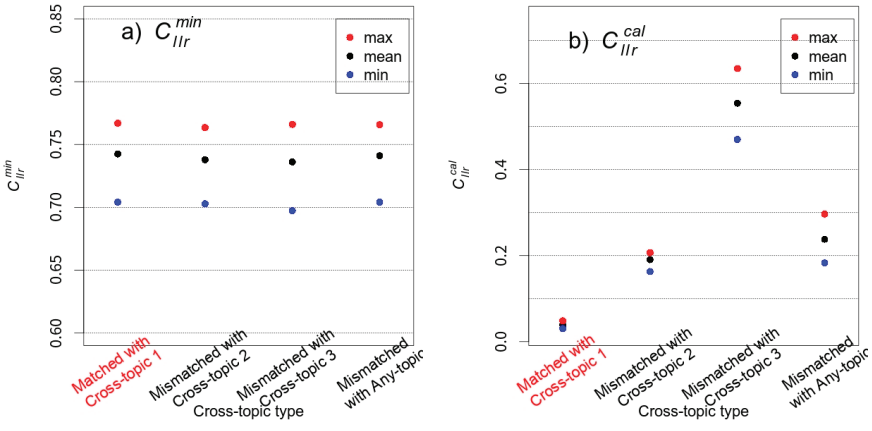


Figure 11. The maximum (max), mean, and minimum (min) C_{llr}^{min} (Panel **a**) and C_{llr}^{cal} (Panel **b**) values of the six cross-validated experiments are plotted for each of the four experimental conditions specified in Table 4.

Panel (b) of Figure 11, which presents the C_{llr}^{cal} values for the four experimental conditions, undoubtedly shows that not using the relevant data considerably impairs the calibration performance. It is interesting to see that the variability of the calibration performance is far smaller for the matched experiment (Cross-topic 1) than for the other mismatched experiments. Note that the maximum, mean, and minimum C_{llr}^{cal} values are very close to each other for the matched experiment (Cross-topic 1). This means that the use of the relevant data is also beneficial in terms of the stability of the calibration performance.

The Tippett plots included in Figure 12 are for the LRs of the four experimental conditions described in Table 4. Note that the LRs of the six cross-validated experiments are pooled together for Figure 12. The deterioration in calibration described for Figure 11b can be visually observed from Figure 12.

Tippett plots, which are also called empirical cumulative probability distributions, show the magnitude of the derived LRs simultaneously for the same-source (e.g., SA) and different-source (e.g., DA) comparisons. In Tippett plots (see Figure 12), the y-axis values of the red curves give the proportion of SA comparisons with $\log_{10}$LR values smaller than or equal to the corresponding value on the x-axis. The y-axis values of the blue curves give the proportion of DA comparisons with $\log_{10}$LR values bigger than or equal to the corresponding value on the x-axis. Generally speaking, a Tippett plot in which the two curves are further apart and in which the crossing-point of the two curves is lower signifies a better performance. Provided that the system is well-calibrated, the LRs above the intersection of the two curves are consistent-with-fact LRs and the LRs below the intersection are contrary-to-fact LRs. In general, the greater the consistent-with-fact LRs are, the better, whereas the smaller the contrary-to-fact LRs are, the better.

The high C_{llr}^{cal} values of the mismatched experiments with Cross-topic 2 (mean C_{llr}^{cal} = 0.19037), Cross-topic 3 (mean C_{llr}^{cal} = 0.55395), and Any-topic (mean C_{llr}^{cal} = 0.23789) show that the resultant LRs are not well-calibrated. The crossing-points of the two curves given in Figure 12b–d, (see the arrows given in Figure 12) deviate from the neutral value of $\log_{10}$LR = 0, further demonstrating poor calibration.

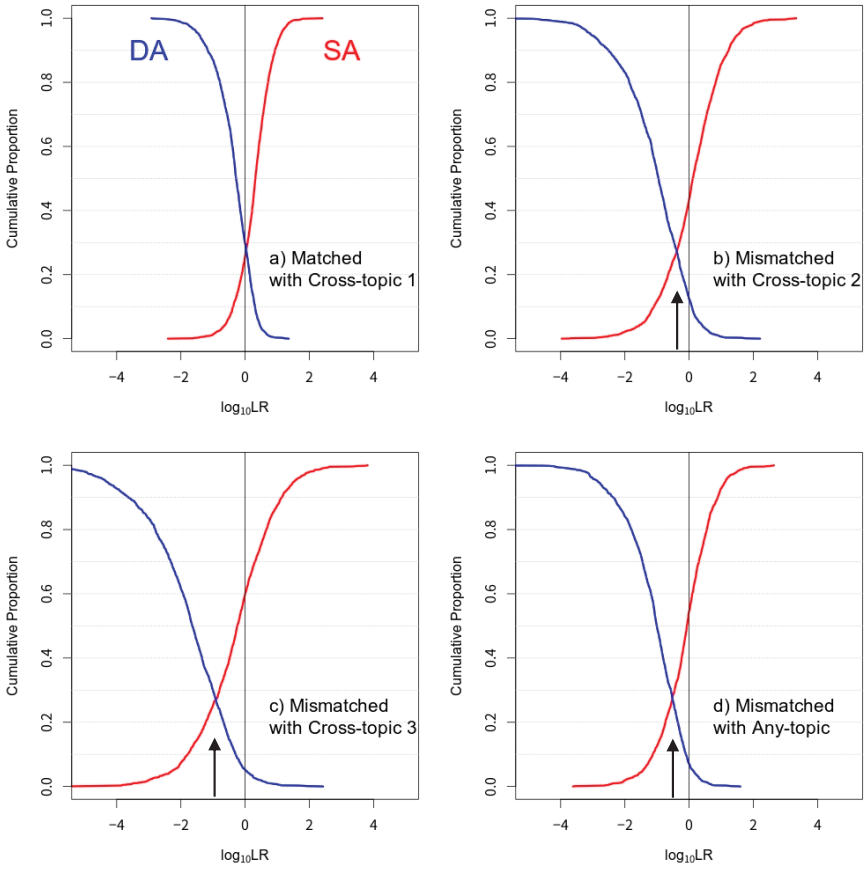


Figure 12. Tippett plots of the LRs derived for the four experimental conditions specified in Table 4. Red curves—SA $\log_{10}$LRs; blue curves—DA $\log_{10}$LRs. Arrows indicate that the crossing-point of the two curves is not aligned with unity. Note that some $\log_{10}$LR values go beyond the range given in the x-axis.

The consistent-with-fact LRs are conservative in magnitude for the matched experiment with Cross-topic 1 (see Figure 12a), keeping the magnitude approximately within $\log_{10}\text{LR} = \pm 3$. The magnitude of the contrary-to-fact LRs is also constrained approximately within $\log_{10}\text{LR} = \pm 2$; this is a good outcome. In the mismatched experiments, although the magnitude of the consistent-with-fact LRs is greater than that of the matched experiment, the magnitude of the contrary-to-fact LRs is also unfavorably enhanced (see Figure 12b–d). That is, the LRs derived with irrelevant data (see Figure 12b–d) are at great risk of being overestimated. This overestimation can be exacerbated if the system is not calibrated (see Figure 12b–d).

Figure 11 indicates that the deterioration in overall performance (measured in C_{llr}) is mainly due to the deterioration in calibration performance (measured in C_{llr}^{cal}), and that using irrelevant data, i.e., Cross-topics 2 and 3 and Any-topic in the Reference database, has minimal bearing on the discrimination performance.

Using simulated FTC data, Ishihara (2020) showed that performance degradation/deterioration caused by the limitation of available data was mainly attributed to poor calibration rather than to the poor discriminability potential; near-optimal discrimination performance can be achieved with samples from as few as 40–60 authors. Furthermore, in his forensic voice comparison (FVC) study investigating the impact of sample size on the performance of an FVC system, Hughes (2017) reported that the system performance was

most sensitive to the number of speakers included in the Test and Calibration databases. The performance was not particularly influenced by the number of reference speakers. Although Ishihara's and Hughes' studies focus on the amount of data as a factor in the system performance, more specifically, the number of sources from which samples are collected, their results equally indicate that the calibration performance is more sensitive to the sample size than is the discrimination performance.

In the current study, the quantity of the data included in each database is sizable: 592 SA and 592 DA comparison for each experiment. Thus, unlike for Ishihara (2020) and Hughes (2017), the degraded aspect of data in the present study is not the quantity but the quality; namely, the degree of topic mismatch between Test and Reference/Calibration databases. It is conjectured that any adverse conditions in data, quantity, or quality, tend to do more harm on the calibration performance than the discrimination performance. However, this requires further investigation.

7. Summary and Discussion

Focusing on the mismatch in topics between the source-questioned and source-known documents, the present study showed how the trier-of-fact could be misled if the validation were NOT carried out:

- under conditions reflecting those of the case under investigation, and
- using data relevant to the case.

This study empirically demonstrated that the importance of the above requirements for validation is true for FTC.[10]

Although the necessity of validation for the admissibility of authorship evidence in court is well acknowledged in the community (Ainsworth and Juola 2019; Grant 2022; Juola 2021), to the best of our knowledge, the importance of the above requirements has never been explicitly stated in relevant authorship studies. This may be because it is rather obvious. However, we would like to emphasize the importance of the above validation requirements in this paper because forensic practitioners may think that they need to use heterogenous corpora in order to make up for the lack of specific corpora; for example, not having enough time to create a customized one, or thinking that the validation of any source-inference systems should be conducted by simultaneously covering a wide variety of conditions; for example, various types of mismatches should be considered. The inclusion of diverse conditions for validation is assumedly a legitimate way of understanding how well the system generally works. However, it does not necessarily mean that the same system works equally well for each specific situation; i.e., the unique condition of a given casework.

If one is working on a case in which the authorship of a given text is disputed and it is a hand-written text, the forensic expert would surely not use social media texts to validate the system with which the authorship analysis is performed. Likewise, they would not use the social media samples as the Reference and Calibration databases in order to calculate an LR for the hand-written text evidence. This analogy goes beyond the use of the same medium for validation and applies to various factors that influence one's own way of writing.

This study focused on the mismatch in topics as a case study to demonstrate the importance of validation. Topic is a vague term, and the concept is not necessarily categorical; thus, it is a challenging task to classify documents into different topics/genres. One document may consist of multiple topics and each topic may be composed of multiple sub-topics. Making matters worse, as pointed out in Section 1.3, topic is only one of many other factors that possibly shape individuals' writing styles. Thus, in real casework, the level of mismatch between the documents to be compared is highly variable and case specific, and databases replicating the case conditions may need to be built from scratch if suitable sources are not available. As such, it is sensible to ask what casework conditions need to be rigorously considered during validation and what other conditions can be overlooked, and these questions need to be pursued in the relevant academic community. These questions are inexorably related to the meaning of relevance. What are the relevant data (e.g., same/similar topics and medium) and relevant population (e.g., non-native use

of a language; same assumed sex as the offender for some languages) (Hicks et al. 2017; Hughes and Foulkes 2015; Morrison et al. 2016)?

Computational authorship analysis has made huge progress over the last decade, and related work demonstrated that some sources of variability can be tolerated to a good extent by the systems compared to a decade ago. As the technology advances, fewer factors may become relevant to consider for validation. Authorship analysis can never be performed under perfectly controlled conditions because two documents are never composed under the exact same settings. Despite this inherent difficulty, authorship analysis has been successful. This leads to the conjecture that some external factors that are considered to be sources of variability can be well suppressed by the systems or that the magnitude of the impact caused by these factors may not be as substantial as feared in some cases.

Nevertheless, it is clear that the community of forensic authorship analysis needs to collaboratively attend to the issues surrounding validation, and to come up with a consensus, perhaps in the form of validation protocols or guidelines, regardless of the FTC approaches to be used. Although it is impossible to avoid some subjective judgement regarding the sufficiency of the reflectiveness of the casework conditions and the representativeness of the data relevant to the case (Morrison et al. 2021), validation guidelines and protocols should be prepared following the results of empirical studies. In fact, we are in a good position in this regard as there are already some guidelines and protocols for us to learn from; some of them are generic (Willis et al. 2015), and others are area-specific (Drygajlo et al. 2015; Morrison et al. 2021; Ramos et al. 2017) or approach-specific (Meuwly et al. 2017).

There are some possible ways of dealing with the issues surrounding the mismatches. One is to look for stylometric features that are robust to the mismatches (Halvani et al. 2020; Menon and Choi 2011), for example limiting the features to those that are claimed to be topic-agnostic (Halvani and Graner 2021; Halvani et al. 2020). Another is to build statistical models that can predict and compensate for the issues arising from the mismatches (Daumé 2009; Daumé and Marcu 2006; Kestemont et al. 2018).

Besides these approaches, an engineering approach is assumed to be possible; e.g., the relevant data are algorithmically selected and compiled considering the similarities to the source-questioned and source-known documents (Morrison et al. 2012) or they may even be synthesized using text-generation technologies (Brown et al. 2020). Nevertheless, these demand further empirical explorations.

The present study only considered one statistical model (Ishihara 2023) but there are other algorithms that might be more robust to mismatches, for example, methods designed for authorship verification that contain random variations in their algorithms (Kocher and Savoy 2017; Koppel and Schler 2004). Another avenue of future study is the application of a deep-learning approach to FTC. A preliminary LR-based FTC study using stylistic embedding reported promising results (Ishihara et al. 2022).

As briefly mentioned above, applying validation to FTC in a manner that reflects the casework conditions and uses relevant data most likely requires it to be performed independently for each case because each case is unique. This further necessitates custom-collected data for each casework. Given this need, unless an appropriate database already exists, the sample size—vis-à-vis both the length of a document and the number of authors documents are collected from—is an immediate issue as it is unlikely to be possible to collect an appropriate amount of data due to various constraints in a forensically realistic scenario. System performance is sensitive to insufficient data, in particular the number of sources from which samples are being collected, both in terms of its accuracy and reliability (Hughes 2017; Ishihara 2020). Thus, extended work is also required to assess the potential tradeoffs between the robustness of FTC systems and the data size,[11] given the limitations of time and resources in FTC casework. Fully Bayesian methods whereby the LRs are subject to shrinkage depending on the degree of uncertainty (Brümmer and Swart 2014) would be a possible solution to the issues of sample size. That is, following Bayesian logic, the LR value should be closer to unity with smaller samples as the uncertainty will be higher.

8. Conclusions

This paper endeavored to demonstrate the application of validation procedures in FTC, in line with the general requirements stipulated in forensic science more broadly. By doing so, this study also highlighted some crucial issues and challenges unique to textual evidence while deliberating on some possible avenues for solutions to these. Any research on these issues and challenges will contribute to making a scientifically defensible and demonstrably reliable FTC method available. This will further enable forensic scientists to perform the analysis of text evidence accurately, reliably, and in a legally admissible manner, while improving the transparency and efficacy of legal proceedings. For this, we need to capitalize on the accumulated knowledge and skills in both forensic science and forensic linguistics.

Author Contributions: Conceptualization, S.I., S.K., M.C., S.E. and A.N.; methodology, S.I., S.K., M.C., S.E. and A.N.; software, S.I. and S.K.; validation, S.I. and S.K.; formal analysis, S.I. and S.K.; investigation, S.I., S.K., M.C., S.E. and A.N.; resources, S.I., S.K. and M.C.; data curation, S.I. and S.K.; writing—original draft preparation, S.I.; writing—review and editing, S.I., S.K., M.C., S.E. and A.N.; visualization, S.I. and S.K.; supervision, S.I.; project administration, S.I.; funding acquisition, S.I., S.K. and M.C. All authors have read and agreed to the published version of the manuscript.

Funding: The contributions from Shunichi Ishihara, Sonia Kulkarni, and Michael Carne were partially supported by an anonymous institution that prefers not to disclose its identity.

Data Availability Statement: The numerical version of the data and the codes (R/Python) used in this study are available from the corresponding author.

Acknowledgments: The authors thank the reviewers for their useful comments.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Table A1. The 140 tokens used for the bag-of-words model.

.	the	“ (open)	” (close)	I	and	a
to	it	of	is	for	that	in
this	you	with	my	on	have	but
n't	's	not	are	was	as	The
)	be	It	(	so	or	like
!	one	do	can	they	use	very
at	just	all	This	out	has	up
from	would	more	good	your	if	an
me	when	' '	these	had	them	will
“	than	about	get	great	does	well
really	product	which	other	some	did	no
time	-	much	've	only	also	little
:	'm	because	there	used	by	what
too	been	any	even	easy	using	am
. . .	were	we	better	could	work	after
into	nice	first	make	over	off	love
need	They	how	still	two	think	?
then	If	price	way	bit	My	who
back	want	their	quality	most	works	made
find	years	see	few	enough	long	now

The formula for calculating a score for the source-questioned ($X = \{x_1, x_2, \cdots, x_k\}$) and source-known ($Y = \{y_1, y_2, \cdots, y_k\}$) documents with the Dirichlet-multinomial model is given in Equation (A1), in which $B(\cdot)$ is a multinomial beta function and $A = \{\alpha_1, \alpha_2, \cdots, \alpha_k\}$ is a parameter set for the Dirichlet distribution. The index k is 140.

$$\text{score} = \frac{B(A)B(A+X+Y)}{B(A+X)B(A+Y)} \tag{A1}$$

The parameters of the Dirichlet model ($A = \{\alpha_1, \alpha_2, \cdots, \alpha_k\}$) were estimated using the Reference database with the maximum likelihood estimation. A derivational process from Equation (1) to Equation (A1) is explicated in Ishihara (2023).

Equation (A2) is for calculating C_{llr}.

$$C_{llr} = \frac{1}{2}\left(\frac{1}{N_{SA}}\sum_{i}^{N_{SA}}\log_2\left(1+\frac{1}{LR_{SA_i}}\right) + \frac{1}{N_{DA}}\sum_{j}^{N_{DA}}\log_2\left(1+LR_{DA_j}\right)\right) \tag{A2}$$

In Equation (A2), LR_{SA_i} and LR_{DA_j} are the linear LR values corresponding to SA and DA comparisons, respectively, and N_{SA} and N_{DA} are the numbers of the SA and DA comparisons, respectively.

For example, linear LR values of 10 and 100 for DA comparisons are contrary-to-fact LR values. The latter strongly supports the contrary hypothesis more than the former; thus, the latter should be more severely penalized than the former in terms of C_{llr}. In fact, the cost for the latter, 6.65821 ($= log_2(1+100)$), is higher than that for the former, 3.4534 ($= log_2(1+10)$).

Notes

1 There are various types of forensic evidence, such as DNA, fingerprints, and voice analysis. The corresponding verification systems demonstrate varying degrees of accuracy for each. Authorship evidence is likely to be considered less accurate compared to other types within the biometric menagerie (Doddington et al. 1998; Yager and Dunstone 2008).

2 There is an argument suggesting that these requirements may not be uniformly applicable to all forensic-analysis methods with equal success (Kirchhüebel et al. 2023). It is proposed that a customized approach to method validation is necessary, contingent upon the specific analysis methods.

3 https://pan.webis.de/clef19/pan19-web/authorship-attribution.html (accessed on 3 February 2021).

4 Instead of more common terms such as 'forensic authorship attribution', 'forensic authorship verification', and 'forensic authorship analysis', the term 'forensic text comparison' is used in this study. This is to emphasize that the task of the forensic scientist is to compare the texts concerned and calculate an LR for them in order to assist the trier-of-fact's decision on the case.

5 T-SNE is a statistical method for mapping high-dimensional data to a two- or three-dimensional space. It was performed with the T-SNE function of Python 'sklearn' library with 'random_state = 123' and 'perplexity = 50'.

6 More specifically 'bert-base-uncased' was used as the pre-trained model with 'max_position_embedding = 1024'; 'max_length = 1024'; and 'padding = max_length'.

7 T-SNE is non-deterministic. Therefore, the T-SNE plots were generated multiple times, both with and without normalizing the document number. However, the result is essentially the same regardless of the normalization.

8 If the output of the Dirichlet-multinomial system is well-calibrated, it is an LR, not a score. Thus, it does not need to be converted to an LR at the calibration stage.

9 This is true as long as the LR is greater than zero and smaller than infinity.

10 It is important to note that the present paper covers only the validation of FTC systems or systems based on quantitative measurements. There are other forms of validation when not quantifying features (Mayring 2020).

11 Some authors of the present paper, who are also FTC caseworkers, are often given a large amount of texts written by the defendant for FTC analyses. Thus, the amount of data in today's cases could be huge, leading to the opposite problem of having too much data. However, when it comes to the data for the use of validation, e.g., Test, Reference, and Calibration data, it could still be a challenging task to collect an adequate amount of data from a sufficient number of authors.

References

Ainsworth, Janet, and Patrick Juola. 2019. Who wrote this: Modern forensic authorship analysis as a model for valid forensic science. *Washington University Law Review* 96: 1159–89.

Aitken, Colin, and Franco Taroni. 2004. *Statistics and the Evaluation of Evidence for Forensic Scientists*, 2nd ed. Chichester: John Wiley & Sons.

Aitken, Colin, Paul Roberts, and Graham Jackson. 2010. *Fundamentals of Probability and Statistical Evidence in Criminal Proceedings: Guidance for Judges, Lawyers, Forensic Scientists and Expert Witnesses*. London: Royal Statistical Society. Available online: http://www.rss.org.uk/Images/PDF/influencing-change/rss-fundamentals-probability-statistical-evidence.pdf (accessed on 4 July 2011).

Association of Forensic Science Providers. 2009. Standards for the formulation of evaluative forensic science expert opinion. *Science & Justice* 49: 161–64. [CrossRef]

Ballantyne, Kaye, Joanna Bunford, Bryan Found, David Neville, Duncan Taylor, Gerhard Wevers, and Dean Catoggio. 2017. An Introductory Guide to Evaluative Reporting. Available online: https://www.anzpaa.org.au/forensic-science/our-work/projects/evaluative-reporting (accessed on 26 January 2022).

Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, and Akitaka Matsuo. 2018. quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software* 3: 774. [CrossRef]

Boenninghoff, Benedikt, Steffen Hessler, Dorothea Kolossa, and Robert Nickel. 2019. Explainable authorship verification in social media via attention-based similarity learning. Paper presented at 2019 IEEE International Conference on Big Data, Los Angeles, CA, USA, December 9–12.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems* 33: 1877–901.

Brümmer, Niko, and Albert Swart. 2014. Bayesian calibration for forensic evidence reporting. Paper presented at Interspeech 2014, Singapore, September 14–18.

Brümmer, Niko, and Johan du Preez. 2006. Application-independent evaluation of speaker detection. *Computer Speech and Language* 20: 230–75. [CrossRef]

Coulthard, Malcolm, Alison Johnson, and David Wright. 2017. *An Introduction to Forensic Linguistics: Language in Evidence*, 2nd ed. Abingdon and Oxon: Routledge.

Coulthard, Malcolm, and Alison Johnson. 2010. *The Routledge Handbook of Forensic Linguistics*. Milton Park, Abingdon and Oxon: Routledge.

Daumé, Hal, III. 2009. Frustratingly easy domain adaptation. *arXiv* arXiv:0907.1815. [CrossRef]

Daumé, Hal, III, and Daniel Marcu. 2006. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research* 26: 101–26. [CrossRef]

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. Paper presented at 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, June 9–12.

Doddington, George, Walter Liggett, Alvin Martin, Mark Przybocki, and Douglas Reynolds. 1998. SHEEP, GOATS, LAMBS and WOLVES: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation. Paper presented at the 5th International Conference on Spoken Language Processing, Sydney, Australia, November 30–December 4.

Drygajlo, Andrzej, Michael Jessen, Sefan Gfroerer, Isolde Wagner, Jos Vermeulen, and Tuija Niemi. 2015. Methodological Guidelines for Best Practice in Forensic Semiautomatic and Automatic Speaker Recognition (3866764421). Available online: http://enfsi.eu/wp-content/uploads/2016/09/guidelines_fasr_and_fsasr_0.pdf (accessed on 28 December 2016).

Evett, Ian, Graham Jackson, J. A. Lambert, and S. McCrossan. 2000. The impact of the principles of evidence interpretation on the structure and content of statements. *Science & Justice* 40: 233–39. [CrossRef]

Forensic Science Regulator. 2021. Forensic Science Regulator Codes of Practice and Conduct Development of Evaluative Opinions. Available online: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/960051/FSR-C-118_Interpretation_Appendix_Issue_1__002_.pdf (accessed on 18 March 2022).

Good, Irving. 1991. Weight of evidence and the Bayesian likelihood ratio. In *The Use of Statistics in Forensic Science*. Edited by Colin Aitken and David Stoney. Chichester: Ellis Horwood, pp. 85–106.

Grant, Tim. 2007. Quantifying evidence in forensic authorship analysis. *International Journal of Speech, Language and the Law* 14: 1–25. [CrossRef]

Grant, Tim. 2010. Text messaging forensics: Txt 4n6: Idiolect free authorship analysis? In *The Routledge Handbook of Forensic Linguistics*. Edited by Malcolm Coulthard and Alison Johnso. Milton Park, Abingdon and Oxon: Routledge, pp. 508–22.

Grant, Tim. 2022. *The Idea of Progress in Forensic Authorship Analysis*. Cambridge: Cambridge University Press.

Halvani, Oren, and Lukas Graner. 2021. POSNoise: An effective countermeasure against topic biases in authorship analysis. Paper presented at the 16th International Conference on Availability, Reliability and Security, Vienna, Austria, August 17–20.

Halvani, Oren, Christian Winter, and Lukas Graner. 2017. Authorship verification based on compression-models. *arXiv* arXiv:1706.00516. [CrossRef]

Halvani, Oren, Lukas Graner, and Roey Regev. 2020. Cross-Domain Authorship Verification Based on Topic Agnostic Features. Paper presented at CLEF (Working Notes), Thessa-loniki, Greece, September 22–25.
Hicks, Tacha, Alex Biedermann, Jan de Koeijer, Franco Taroni, Christophe Champod, and Ian Evett. 2017. Reply to Morrison et al. (2016) Refining the relevant population in forensic voice comparison—A response to Hicks et al. ii (2015) The importance of distinguishing information from evidence/observations when formulating propositions. *Science & Justice* 57: 401–2. [CrossRef]
Hughes, Vincent. 2017. Sample size and the multivariate kernel density likelihood ratio: How many speakers are enough? *Speech Communication* 94: 15–29. [CrossRef]
Hughes, Vincent, and Paul Foulkes. 2015. The relevant population in forensic voice comparison: Effects of varying delimitations of social class and age. *Speech Communication* 66: 218–30. [CrossRef]
Ishihara, Shunichi. 2017. Strength of linguistic text evidence: A fused forensic text comparison system. *Forensic Science International* 278: 184–97. [CrossRef] [PubMed]
Ishihara, Shunichi. 2020. The influence of background data size on the performance of a score-based likelihood ratio system: A case of forensic text comparison. Paper presented at the 18th Workshop of the Australasian Language Technology Association, Online, January 14–15.
Ishihara, Shunichi. 2021. Score-based likelihood ratios for linguistic text evidence with a bag-of-words model. *Forensic Science International* 327: 110980. [CrossRef] [PubMed]
Ishihara, Shunichi. 2023. Weight of Authorship Evidence with Multiple Categories of Stylometric Features: A Multinomial-Based Discrete Model. *Science & Justice* 63: 181–99. [CrossRef]
Ishihara, Shunichi, and Michael Carne. 2022. Likelihood ratio estimation for authorship text evidence: An empirical comparison of score- and feature-based methods. *Forensic Science International* 334: 111268. [CrossRef] [PubMed]
Ishihara, Shunichi, Satoru Tsuge, Mitsuyuki Inaba, and Wataru Zaitsu. 2022. Estimating the strength of authorship evidence with a deep-learning-based approach. Paper presented at the 20th Annual Workshop of the Australasian Language Technology Association, Adelaide, Australia, December 14–16.
Juola, Patrick. 2021. Verifying authorship for forensic purposes: A computational protocol and its validation. *Forensic Science International* 325: 110824. [CrossRef]
Kafadar, Karen, Hal Stern, Maria Cuellar, James Curran, Mark Lancaster, Cedric Neumann, Christopher Saunders, Bruce Weir, and Sandy Zabell. 2019. American Statistical Association Position on Statistical Statements for Forensic Evidence. Available online: https://www.amstat.org/asa/files/pdfs/POL-ForensicScience.pdf (accessed on 5 May 2022).
Kestemont, Mike, Enrique Manjavacas, Ilia Markov, Janek Bevendorff, Matti Wiegmann, Efstathios Stamatatos, Martin Potthast, and Benno Stein. 2020. Overview of the cross-domain authorship verification task at PAN 2020. Paper presented at the CLEF 2020 Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 9–12.
Kestemont, Mike, Enrique Manjavacas, Ilia Markov, Janek Bevendorff, Matti Wiegmann, Efstathios Stamatatos, Martin Potthast, and Benno Stein. 2021. Overview of the cross-domain authorship verification task at PAN 2021. Paper presented at the CLEF 2021 Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21–24.
Kestemont, Mike, Michael Tschuggnall, Efstathios Stamatatos, Walter Daelemans, Günther Specht, Benno Stein, and Martin Potthast. 2018. Overview of the author identification task at PAN-2018: Cross-domain authorship attribution and style change detection. Paper presented at the CLEF 2018 Conference and the Labs of the Evaluation Forum, Avignon, France, September 10–14.
Kirchhüebel, Christin, Georgina Brown, and Paul Foulkes. 2023. What does method validation look like for forensic voice comparison by a human expert? *Science & Justice* 63: 251–57. [CrossRef]
Kocher, Mirco, and Jacques Savoy. 2017. A simple and efficient algorithm for authorship verification. *Journal of the Association for Information Science and Technology* 68: 259–69. [CrossRef]
Koppel, Moshe, and Jonathan Schler. 2004. Authorship verification as a one-class classification problem. Paper presented at the 21st International Conference on Machine Learning, Banff, AB, Canada, July 4–8.
Koppel, Moshe, Shlomo Argamon, and Anat Rachel Shimoni. 2002. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing* 17: 401–12. [CrossRef]
López-Monroy, Pastor, Manuel Montes-y-Gómez, Hugo Jair Escalante, Luis Villasenor-Pineda, and Efstathios Stamatatos. 2015. Discriminative subprofile-specific representations for author profiling in social media. *Knowledge-Based Systems* 89: 134–47. [CrossRef]
Lynch, Michael, and Ruth McNally. 2003. "Science", "common sense", and DNA evidence: A legal controversy about the public understanding of science. *Public Understanding of Science* 12: 83–103. [CrossRef]
Mayring, Philipp. 2020. *Qualitative Content Analysis: Theoretical Foundation, Basic Procedures and Software Solution*. Klagenfurt: Springer.
McMenamin, Gerald. 2001. Style markers in authorship studies. *International Journal of Speech, Language and the Law* 8: 93–97. [CrossRef]
McMenamin, Gerald. 2002. *Forensic Linguistics: Advances in Forensic Stylistics*. Boca Raton: CRC Press.
Menon, Rohith, and Yejin Choi. 2011. Domain independent authorship attribution without domain adaptation. Paper presented at International Conference Recent Advances in Natural Language Processing 2011, Hissar, Bulgaria, September 12–14.
Meuwly, Didier, Daniel Ramos, and Rudolf Haraksim. 2017. A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation. *Forensic Science International* 276: 142–53. [CrossRef]
Morrison, Geoffrey. 2011. Measuring the validity and reliability of forensic likelihood-ratio systems. *Science & Justice* 51: 91–98. [CrossRef]

Morrison, Geoffrey. 2013. Tutorial on logistic-regression calibration and fusion: Converting a score to a likelihood ratio. *Australian Journal of Forensic Sciences* 45: 173–97. [CrossRef]
Morrison, Geoffrey. 2014. Distinguishing between forensic science and forensic pseudoscience: Testing of validity and reliability, and approaches to forensic voice comparison. *Science & Justice* 54: 245–56. [CrossRef]
Morrison, Geoffrey. 2018. The impact in forensic voice comparison of lack of calibration and of mismatched conditions between the known-speaker recording and the relevant-population sample recordings. *Forensic Science International* 283: E1–E7. [CrossRef]
Morrison, Geoffrey. 2022. Advancing a paradigm shift in evaluation of forensic evidence: The rise of forensic data science. *Forensic Science International: Synergy* 5: 100270. [CrossRef]
Morrison, Geoffrey, Ewald Enzinger, and Cuiling Zhang. 2016. Refining the relevant population in forensic voice comparison—A response to Hicks et al.ii (2015) The importance of distinguishing information from evidence/observations when formulating propositions. *Science & Justice* 56: 492–97. [CrossRef]
Morrison, Geoffrey, Ewald Enzinger, Vincent Hughes, Michael Jessen, Didier Meuwly, Cedric Neumann, Sigrid Planting, William Thompson, David van der Vloed, Rolf Ypma, and et al. 2021. Consensus on validation of forensic voice comparison. *Science & Justice* 61: 299–309. [CrossRef]
Morrison, Geoffrey, Felipe Ochoa, and Tharmarajah Thiruvaran. 2012. Database selection for forensic voice comparison. Paper presented at Odyssey 2012, Singapore, June 25–28.
Murthy, Dhiraj, Sawyer Bowman, Alexander Gross, and Marisa McGarry. 2015. Do we Tweet differently from our mobile devices? A study of language differences on mobile and web-based Twitter platforms. *Journal of Communication* 65: 816–37. [CrossRef]
Nini, A. 2023. *A Theory of Linguistic Individuality for Authorship Analysis*. Cambridge: Cambridge University Press.
President's Council of Advisors on Science and Technology (U.S.). 2016. *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods*. Available online: https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf (accessed on 3 March 2017).
Ramos, Daniel, and Joaquin Gonzalez-Rodriguez. 2013. Reliable support: Measuring calibration of likelihood ratios. *Forensic Science International* 230: 156–69. [CrossRef] [PubMed]
Ramos, Daniel, Juan Maroñas, and Jose Almirall. 2021. Improving calibration of forensic glass comparisons by considering uncertainty in feature-based elemental data. *Chemometrics and Intelligent Laboratory Systems* 217: 104399. [CrossRef]
Ramos, Daniel, Rudolf Haraksim, and Didier Meuwly. 2017. Likelihood ratio data to report the validation of a forensic fingerprint evaluation method. *Data Brief* 10: 75–92. [CrossRef]
Rivera-Soto, Rafael, Olivia Miano, Juanita Ordonez, Barry Chen, Aleem Khan, Marcus Bishop, and Nicholas Andrews. 2021. Learning universal authorship representations. Paper presented at the 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic, April 17.
Robertson, Bernard, Anthony Vignaux, and Charles Berger. 2016. *Interpreting Evidence: Evaluating Forensic Science in the Courtroom*, 2nd ed. Chichester: Wiley.
Stamatatos, Efstathios. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology* 60: 538–56. [CrossRef]
van der Maaten, Laurens, and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9: 2579–605.
van Leeuwen, David, and Niko Brümmer. 2007. An introduction to application-independent evaluation of speaker recognition systems. In *Speaker Classification I: Fundamentals, Features, and Methods*. Edited by Christian Müller. Berlin/Heidelberg: Springer, pp. 330–53.
Willis, Sheila, Louise McKenna, Sean McDermott, Geraldine O'Donell, Aurélie Barrett, Birgitta Rasmusson, Tobias Höglund, Anders Nordgaard, Charles Berger, Marjan Sjerps, and et al. 2015. Strengthening the Evaluation of Forensic Results Across Europe (STEOFRAE): ENFSI Guideline for Evaluative Reporting in Forensic Science. Available online: http://enfsi.eu/wp-content/uploads/2016/09/m1_guideline.pdf (accessed on 28 December 2018).
Yager, Neil, and Ted Dunstone. 2008. The biometric menagerie. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32: 220–30. [CrossRef]
Zhang, Chunxia, Xindong Wu, Zhendong Niu, and Wei Ding. 2014. Authorship identification from unstructured texts. *Knowledge-Based Systems* 66: 99–111. [CrossRef]

languages

MDPI

Article

On Genre as the Primary Unit of Language (Not Only) in Law

Dieter Stein

English Language and Linguistics Department, Heinrich-Heine-Universität Düsseldorf, 40225 Düsseldorf, Germany; stein@hhu.de

Abstract: Taking as points of departure modern pragmatic theory and the information-theoretic view of communication offered by Levinson, this paper re-defines the notion of "genre" as a primarily top-down functioning kind of pre-existing, conventionalized package deal in construing meaning. As a consequence, this paper argues for relativizing the role of code (langue), given information in favor of pre-existing pragmatic-functional knowledge in "making meaning". This discussion is focused on law, which is where the issue of whether and how much meaning is "in the text", and what it means to be "in the text" is theoretically and practically paramount.

Keywords: text type; genre; cognition; psycholinguistics; macrostructure; superstructure; context; focus; law; interpretation; meaning; domain; discourse

1. Scope of the Paper: Between Law, Discourse, and Information Theory

How can we know so much, given that we know so little?
How can we know so little, given that we know so much?

This paper tries to deal with theoretical concerns raised by the near-ubiquitous notion of "genre", with a special, but not exclusive, interest in its application to law.[1] Being a functionally, language-externally and socially motivated phenomenon, discussions of the concept tend to be triggered by language—external developments, such as the rise of new media (Giltrow and Stein 2009; Anesa and Engberg 2023), generating new demands on knowledge management in new domains. Law, although intimately related to language, is such an external societal, cultural phenomenon, organized in genres. Law is an institutional and societal context characterized, if not defined, by a functionally coherent, but linguistically diverse landscape of genres. As a theoretically oriented paper, the concern here is not with a practical, functional analysis of specific genres, but it may have bearing on how and why such analyses are conducted.

The second source of theoretical interest in the concept of genre is found in Levinson's recent paper on the "dark matter" of pragmatics (Levinson 2024). Levinson argues that communication—especially human, linguistic communication—needed to develop strategies for coping with informational "bottlenecks", in particular, ones that enforce affordable strategies of redundancy to enable communication to take place in a narrow tunnel of capacity and time. I will argue that the concept of genre is one such strategy, evolutionarily evolved, and which is part of what Levinson would call the "known dark matter" in communication (On the other side of "known dark matter" is the "unknown dark matter" of pragmatics: the vast array of strategies we know must exist, but which are yet completely unknown). Among these strategies, the signal carried by linguistic forms has been overvalued, while the role of other, non-verbal information has gone largely unappreciated.

Law is an area that tends to be dominated by a "folk" ideology of language that sees the linguistic surface information as the main source of signal information. So, the third major theoretical angle of approach in this paper is through the ongoing debates about "meaning making" in modern, essentially neo-Gricean pragmatics, with its conception of an extraction or construction of meaning in a staged logical order that essentially involves the adduction of non-linguistically given knowledge at several stages. It is this non-linguistic

Citation: Stein, Dieter. 2024. On Genre as the Primary Unit of Language (Not Only) in Law. *Languages* 9: 333. https://doi.org/10.3390/languages9110333

Academic Editors: Julien Longhi and Nadia Makouar

Received: 25 June 2024
Revised: 14 October 2024
Accepted: 15 October 2024
Published: 25 October 2024

knowledge that is the focus of this paper. It necessarily forms an important issue in law, due to the assumption, prevalent in law, that the content of the law resides within texts. But it must be asked: what is the "text"?

2. Language Occurs in Texts

I claim that the primary unit of the occurrence, not of "language", but of communication involving language, is genre. Any individual, bounded piece of discourse gains its identity as an instantiation of a genre. Genre is a pivotal concept in the linguistic analysis of communication in the domain of law, which is linguistically organized as a complex, interrelated web of genres. The activities within law make continual reference to types of genres in the law, while classifications into genres are a continual topic in the discipline and in discourse at large (Kurzon 1997; Engberg 2013; Taboada 2004). It is therefore at least appropriate, if not obligatory, to look at the theoretical underpinning of the concept of genre from the vantage point of pragmatics.

While it is already well established in practical research, thanks to work by Swales and Bhatia (cf. the survey in Artemeva and Freedman 2016), I will here try to point to more theoretical issues, arguably those with the most salient practical consequences. One battle cry of the earliest "textlinguistic" turn in much of European continental linguistics was that language occurs in texts. This was a consequence of structuralism's undue privileging of the sentence as the base unit of analysis, itself preceded by a focus on the morpheme. There were good reasons, then, for upscaling the explanatory domain of linguistics to a higher unit, because the grammaticality or otherwise of a number of phenomena could not be explained sentence-internally.

There are two competing or complementary concepts of a "text": *Text type*, first, and second, *genre*. Each represents a different perspective on the constitutive object of research, depending on the point of departure. Traditionally, research started from what "occurs" on the "surface" of language. The classic analysis is Biber's famous study (Biber 1995, "Dimensions of Register Variation"). Following a long tradition that explains or "accounts for" surface distributions by classification into registers, where a "functional" explanation of differential surface distribution resorts to a "notional" concept of genres. Biber's perspective on the text "genre" is therefore "interpretive". Genre is called upon as an explanatory dimension for the formal statistical distribution of surface forms. For our purposes, let us refer to the two basic perspectives as "form"-based and function"-based. The former can be seen as bottom-up, and the latter, function-based, as top-down. The former corresponds to "text type" while the latter refers to "genre". In this view, genre is located at the top, and the text (text type) closer to, or even at the bottom, as the linguistic input level.

It is intuitively apparent from the above that the higher on the vertical axis the linguistic unit, the less "autonomous" the unit is: the sources of knowledge tend to come less from language, morphemes, form, or "code" (Lyons' "system sentence", Lyons 1981, p. 196) the "higher" you go in this concept. One issue that has seemed clear for higher units has been a much debated issue lower down, on the level of the sentence and lower: the issue of underdeterminacy. It has always been easier to accept that not only are lower to higher meanings increasingly determined by extra-linguistic information, but also that there are qualitatively different types of meanings ascribed to the global text level that are not explicable as additions of meanings that are, to a stronger extent, componentially generated lower down. So this "basic" type of decrease in autonomy and the later-to-be discussed notions of underdeterminacy are just two sides of the same problem.

What is the relationship between the two perspectives? Where does the researcher start the "explanatory" journey? Top or bottom? Notice that by "direction of explaining" we refer to the meta-decision of where to *logically* start.

Is there a one–one relationship between the two unit types (what I will later call the insulationist view, § 8), as often postulated for the lower end of the scale, such that a

linguistic form would correspond to a higher-level unit? In particular, my question is: can genres be defined linguistically?

A famous dictum of pragmatics is that "form", i.e., text type, "underdetermines" function. Given this assumption, it is not possible to define genres with reference to the occurrence of surface linguistic features. Or, to draw on Giltrow's inverse formulation of the same: it is not possible to define genre by surface linguistic features alone.

If evidence were needed, we could point to Giltrow's paradigmatic analysis of a 1760 "Treaty of Peace and Friendship" between Canadian First Nations and European settlers, where the linguistic material in the text "underdetermines" the genre. In Giltrow's example, the treaty is just a writ, or indeed a contract, from which certain real obligations arise, with concrete political and financial consequences. As the graph postulates, "Form alone" does not determine genre, bottom-up, on the level of the whole discourse.

Viewed top-down, it is well known that a genre like a contract has so many sur—"faces", that is to say many linguistic form possibilities, so many that there is no single way to define it micro-linguistically. Makmillen (2007) offers another example of how, like a chameleon, the same "text" can be read as instantiating several different genres, with drastic political and financial consequences.

3. Genre and Indeterminacy

The concept of genre has a history characterized by an early emphasis on didactic issues—that is, identifying genres so as to make them teachable—and on surface linguistic features, with the concept in close competition with the notion of "register". While persisting in applied pursuits, such as the teaching of writing, the more modern theoretical approach to genre shifts focus to "…a level of genre that represents those [communicative, DS] events which have been culturally recognized" (Taboada 2004, p. 19). An indication of the abstract knowledge nature of the concept of "genre" is also found in Giltrow (2016, p. 220), which relativizes the traditional emphasis of research in the area of genre on linguistic form and which points to the same information-theoretic argument as made by Levinson (2024), which shifts the emphasis to a massive, but as yet underemphasized, share of ineluctable non-linguistic information. The underdetermination of meaning by form is sometimes an efficiency, sometimes a liability, and sometimes a resource.

More broadly, the new genre theory can make a contribution to the work on the pragmatic problem of underdetermination. The gist is the following: spheres of activity constitute common ground, narrowing context to a common ground of accessible assumptions from which interlocutors can, and do, draw inferences. But the contribution comes at a cost for genre theory; for now, the main term to be reckoned with in genre theory is *mutual consciousness*, in all the silent trepidations detected by pragmatics: assumptions unascertainable, uneven, unstable and changeable, even when form remains unchanged (Giltrow 2016, p. 220).

A question, of course, is what the definitional features of a genre are, if not form. Giltrow makes several key points on this question. An acutely felt "cost" for genre theory is the loss, absence, or radical downgrade in the status, of purportedly "objective" surface information. An interactive view of text and genre finds little, if anything, in the morphemes on the "surface" of a text. It thus pits itself against the "insulationist" view, that genre knowledge is autonomous, extractable from code.

It is this fact—the very indirect or minimal access to genre from surface features—that causes such embarrassment and bewilderment, especially for a linguistic perspective that is used to starting from the bottom up. The tendency towards minimalism is understandable, as surface features seem to provide something that is observable, objective, and therefore more intersubjectively recognizable. As such, minimalist approaches are more congenial to legal endeavors. More countable, they are attractive for computational approaches and for lawyers in the same way. If minimalism means the provenience of only minimal knowledge that comes from contextual, language-external sources, maximalism implies the opposite.

A maximalist view of genre entails a cognitive view, as this external knowledge—the common ground—must be assumed to be present somehow in our cognition. Here, we will focus on the top-down perspective, starting with entities that are not directly observable, entities of a more cognitive nature, manifested at the linguistic surface only very partially and often only in traces. Yet, these cognitive entities are the prime explanatory dimension for forms on the linguistic surface.

There seems to be something of a complaint tradition, regretting the fact of underdetermination. While there are cases of underdeterminacies (Witczak-Plisiecka 2009) that create genuine problems for legal interpretation, the general philosophical complaint about the lack of one-to-one correspondence between form and meaning is a type of language ideology that bedevils many applications of language. Certainly, many current folk–linguistic debates and even applied linguistic pursuits, including forensic linguistics, specifically in the shape of the so-called "referential ideology" (Ainsworth 2008), are still hamstrung by their fostering of this false dilemma. Just as it is a "design feature" of language to have few units, a multitude of realizations and expressive means, on the lower levels, so we should expect the same on the level of genres: never a costly one–one relationship between form and function.

The central background fact—to make an information-theoretic observation—is that human speech encoding is relatively very slow; the actual process of phonetic articulation is a bottleneck in a system that can otherwise run much faster. The speaker is trying to find an economical means of invoking specific ideas in the hearer, knowing that the hearer has exactly this expectation. Now, the solution to the bottleneck is just this: let not only the content, but also the metalinguistic properties of the utterance (its form) carry the message. Or, in other words, find a way to piggyback meaning on top of meaning (cf. Levinson 2000, p. 6).

And emphasizing the necessity to find a redundancy-creating way to minimize the infinite potentialities of extra-linguistic knowledge: "Because these inferences rule out a number of possible states of affairs they multiply the informational load of what has been said by a significant factor" (Levinson 2000, p. 31f).

Genre, then, is much "richer" than the text. Much more content "functions" in a text as discourse and operates on what is there in terms of the text sentence or "what is said" than what is represented at the surface:

1. The highly selective filtering or selection of context or knowledge from all available knowledge: only a fraction of what could be selected is actually—and legitimately—selected to be mutually present in the cognition of communication partners.
2. What there is in terms of surface expressions: the surface expressions themselves have a function within that present knowledge. They are intended by the writer or speaker to produce effects in a cognitively conceived discourse world on the receiver side. It is therefore clear that the content side is a cognitive entity, a view on which I will elaborate on in what follows. To a large extent, *genre is internalized context*, and context is knowledge, as it is represented in our cognition and is assumed by the speaker to be exactly there. However, "knowledge" is potentially everything that could be represented. Restricting all possible knowledge to a relevant and cognitively manageable, efficient amount is the salient function of genre with respect to the information-theoretical issue emphasized by Levinson (2024).

The issue here is that interlocutors mutually assume a tiny fraction of all possible knowledge—"mutual knowledge" by first approximation—to be present. Only the restricted and evoked relative knowledge is present and represents the relevant and legitimate "cognitive environment", where assumed legitimacy is defined by mutual assumptions about exactly which type of genre game is being played, and consequently which knowledge is legitimately assumed to be present at any given point in the interaction. It is an essential component of genre that this element of "legitimacy" of mutual knowledge expectation is present. It is understood that miscalculations about mutual knowledge tend to have a worse effect on communication than incorrect grammar, and therefore, unlike

grammatical mistakes, the preference is always for the immediate repair of mistakes about mutual knowledge. Mutual knowledge is linguistically reflected in linguistic forms that indicate, e.g., definiteness and specificity in the referential-pragmatic sense, as well as in the more general sense of the choice of more or less specific expressions with more or less intentional content. Importantly, that the receiver access the right kind of knowledge is just as much part of the speaker intention as the standard Gricean view of the intentionality of speaker meanings as that which is to be recovered. On a more general level, the co-construction of meaning is now generally recognized (Elder and Jaszczolt 2024, cf. Kuhlen and Rahman 2023 for psycholinguistic aspects).

Two questions now arise that I consider pivotal for my approach to genre. First of all: How can we know so little, given that we know so much? This question takes up the problem of "efficiency" mentioned by Giltrow: reducing the vast possibilities of contextual knowledge, in a manner that enables efficacious comprehension. The complementary version of the question would be the following: How can we know so much, given that we know so little? How is it that we perform such complex knowledge changes and reorderings in our cognitive world, given that these cognitive operations are triggered by such scant linguistic instructions? In terms of Relevance Theory, how can a faint linguistic signal cause in the hearer such massive cognitive effects? From the scantiest linguistic signals, interlocutors strengthen, revise, or abandon available assumptions, even though the linguistically given information is little more than a semi- or un-consciously remembered history of previous communicative encounters. What we have conventionally termed "linguistic meaning" is itself highly constrained by genre.

The issue of the autonomy of a text raises exactly this issue, begging the question: to what extent are these instructions for cognitive re-arrangements ("cognitive effects") due to the "linguistic meaning" of expressions and/or the knowledge types and operation processes on which they operate?

The vulgo version, which keeps appearing in discussions of legal interpretation, of the autonomy issue is: To what extent is the knowledge contained either in the text or in the context? It is clear that the restricting and modulating information comes not only from the cognitive environment but also from the co-occurring referential expressions in the text (long known in literary studies, what literary scholar I.A. Richards colorfully called the "interanimation of words" (Richards [1937] 1964, p. 47)—or, in legal terms: noscitur a sociis):

> According to the insulationist account the meaning of any one word that occurs in a particular sentence is insulated against interference from the meaning of any other word in the same sentence [...] Interactionism makes the contradictory assertion: in some sentences in some languages the meaning of a word in a sentence may be determined in part by the word's verbal context in that sentence. (Recanati 2003, p. 132, citing Cohen)

This co-textual aspect (co-text turned into cognitively present context during the further linear progress of comprehension) is emphasized in its specific shape for legal documents by Skoczen (2016). But the concrete meaning of legal documents is, in turn, itself top-down influenced or massively co-determined by genre and the schemata and knowledge frames called up through it.

Cognitively oriented genre theory therefore holds that genres are bounded, prepackaged types of knowledge, including, but not exhaustively defining linguistic surface expressions, which we can initially characterize here as propositional, componential language knowledge or "linguistic meaning" or "code" knowledge.

It would appear, then, that the highest level of pre-packaged context and comprehension/constructing level is genre. Genre is characterized by two properties.

1. Genres are "a vast association network between utterance forms, contexts and actions" (Levinson 2024, p. 18). Here, "actions", in terms of speech act theory, refers to illocution types. "Utterance forms" are linguistic surface forms, elements of code. The elements of this interpretational process accessed in a way defined by genre are

the following:

"There are various sources of information that contribute to the main meaning conveyed by the speaker and recovered by the addressee. In the DS [default semantics, D.S.] model, we identify five main sources: word meaning and sentence structure (WS), world knowledge (WK), situation of discourse (SD), and two kinds of default information: stereotypes and presumptions about society and culture (SC) and properties of human inferential system (IS). The latter, for example, account for the fact that the strongest, most informative interpretation is the preferred one, such as the referential rather than attributive reading of definite descriptions, de re rather than de dicto reading of propositional attitude reports, or the anaphoric rather than presupposing reading of referential partial matches" (Jaszczolt 2011, p. 21).

2. The meaning total (meaning NN) of a concrete text instance of a genre (a contextually fully interpreted instance of a "text-sentence") is what Jaszczolt (2011) terms the "primary meaning" or the "default meaning". It is arrived at instantaneously and effortlessly; Levinson indeed stresses the instantaneous speed at which "participants actually ascribe actions to utterances" (cf. Levinson 2024, p. 16f). Although Levinson's focus is primarily on spoken conversation, his description of the situation holds for processing texts as members of genres. For spoken discourse, "speech act recognition can occur very early during the processing of an incoming turn—even within the first syllable or two" (Levinson 2024, p. 17). Even though, as in written communication, there are cases when communicants take time to reflect, in spoken situations, there is simply no time to deliberately figure out what action is intended.

   Even so, the cardinal phenomenon persists that utterances are instantaneously processed. It would seem, then, that the choice of "actions" is either extremely limited at any point of processing, or indeed pre-determined, so that no further decisions and only "very intensive cognitive reasoning" (Levinson 2024, p. 18) are required. Integral parts of genres are, in the van Dijk-ian sense, their "superstructures": types and conventionalized functional parts, often in a conventional sequenced order. Classic examples might include the hero's narrative and the argumentative essay. The existence of genre is therefore an essential part of the "efficiency" aspect of processing language. In Kahneman's (2012, p. 20) neuro-psychological metaphor, genre is part of our "System 1". Kahneman writes the following: "*System 1* operates automatically and quickly, with little or no effort and no sense of voluntary control". It is a fully automated, fast way of thinking that, in its default mode of operation, does not require special mental effort. This way of thinking is against "System 2", which takes up significant mental space by requiring special and conscious processing effort. As Kahneman (2012, p. 21) again writes, "System 2 allocates attention to the effortful mental activities that demand it, including complex computations. [. . .] The automatic operations of System 1 generate surprisingly complex patterns of ideas, but only the slower System 2 can construct thoughts in an orderly series of steps".

4. Genre as Abstract Entities

It has been argued, in culturalist, Foucaultian discussions of language and law, that there is an even higher abstract level of genre knowledge. Such discussions treat cultural knowledge as largely packaged in terms of disciplines of knowledge, as organized in societal domains, such as sciences, religion, literature, and law—in their own non-technical sense, as genre.

Legal scholars are well aware that their field is, or can be viewed as, a genre. (FN 1) Law, though it deals with "society" and human experience, involves distinct forms (particular legal texts, specific practices, procedures, language, categories, etc.) that distinguish its manner of representing and processing such experience. The idea that law, as genre, is somehow distinct from, yet related to its surroundings has often been channeled through theoretical claims about the relations between (sometimes only vaguely defined) "law" and "society". These analyses have preoccupied scholars for a long time now, and when they

involve causal claims, they are debates about the level of independence of genres from extra-linguistic determination.

Rosenberg refers to the fundamental legal–internal controversy over whether law is externally conditioned (culturally/sociologically) or is a matter of law-internal rules and dogmatics. While we cannot pursue the question here let us note that the issue—whether genre is a self-contained system or one that is determined externally—is relevant to the question of text-type vs. genre. Can or should we consider texts products of system-internal rules or as genre, with massive external determination and interaction?

Not only is the dichotomy repeated within the domain of law, but the fact that legal genres are—like any genre—determined externally also appears to speak to a non-positivistic stance.

Abstraction away from the level of texts is taken one step further by Rosenberg, in a historical analysis of advertising:

> Law, understood as a dynamic part of cultural negotiation rather than a predefined profession, discipline or institutional setting, became implicated in directing the system of advertising in the different media, defining approaches for analysing and understanding it, and determining its boundaries. Ultimately, by performing boundary work law shaped the status of advertising, which this book shows to have emerged as deeply conflicted: legal means constituted advertising as a legitimate and indeed indispensable system of modern life, but also as a disparaged, ridiculed and criticized one, considered suspect in epistemological and aesthetic terms. (Rosenberg 2022, p. 10)

In Rosenberg's approach, the very domain of law as such is, at its highest level of existence, elevated to an abstract context of process and procedurality, far away from any level of componentiality. But law is, as with all elements of what can define a genre, determinative of top-down processes of meaning building. It is also, in Levinson's conceptual grid, part of the known, not even of the known unknown.

Abstraction of genre from surface forms makes any instantiation of genre much more than the sum of its surface forms. In turn, part of a discourse or discourse types receive their function.

A narrative receives its function, and the rules of the game for handling and interpreting it, from a larger genre configuration (like law) in which it is embedded. The raw, functionally uninterpreted narrative is an ineffectual concept. A literary environment and a judicial environment—or any other number of environments, for that matter—would each make different things out of the same surface narrative. A similar case are acts that only together make sense as a mobbing episode (Stein 2022), both for the analyst as well as for the awareness of the victim, who as a rule do not interpret constituent acts as part of a larger negative communicative scheme.

It is, in addition, a further instantiation of the phenomenon that genres can be organized in hierarchical form, such as is obviously the case in phenomena like harassment, where a partial process of negative actions receive their super-summative function through being part of a larger abstract unit "harassment" (Guillén-Nieto 2024).

We will perhaps cease to refer to this effect as a problem if we take into account what Bhatia has repeatedly referred to as the way the genres in a domain like law (or genre in Rosenberg's sense) are intertextually and interdiscursively appropriated (Bhatia 2023, p. 159).

I will argue, therefore, that the reductional redundancies of knowledge as in System no. 1 are definitional for genre. The question remains: how do we reduce the globally possible knowledge types to a level that is cognitively manageable in terms of mental/psycho-ling processing capacities? Such reduction occurs on all hierarchical levels of discourse, on the level at which knowledge is called upon, as much as on the constraining of which knowledge to invoke at the level of the domain, or—in the Rosenbergian sense—at the level of genre. It would seem we must assume that these knowledge types are variously present and salient to different degrees, as manifested by their surface treatment as definite or not. But it is exactly this mechanism of moving knowledge into local saliency, and only that and

no other knowledge or entities, that is the achievement of genre. So a genre is an n-tuple of very different types of knowledge involving selected types of games and actions with cultural and social knowledge (whether that is, for example, conversation or legal statute) as well as processing strategies. Linguistic expressions, including their meaning and reference potentials, as with specific linguistic meanings, receive their specific functions through top-down processes, even at the earliest stages of proposition creation. Even the first "pre-selections", on the way from system to text sentence, are pragmatically determined and thereby determined genre-wise. Beyond the now well-researched micro-linguistic surface forms, so often studied in connection with Register studies, the following is a selection of—in the widest sense—processing conventions or pragmatic conventions associated with different genres, such as the following: rules of interpretation, presumptions about how to process, discourse rules and conversational rules. It is a specific sub-selection of these knowledge types that defines a genre and constrains our processes of meaning making.

Taking a statute as an illustrative example, part of the package deal of statute as a genre are interpretive principles like the following:

- literalist presumption
- parole evidence rule
- surplusage rule
- four corners rule
- legal canons of interpretation

Walton et al. (2021) described the types of knowledge that are presumed to be called on in resolving a concrete legal case and which provide arguments in the legal case:

1. Structural presumption (a presumption related to the "structure" of a legal text or system) (Argument of economy): "Legal texts normally have no redundant expressions".
2. Linguistic, specific presumption (Systems argument): "By 'control,' the statute means control in the sense that the tenant has permitted access to the premises".
3. Linguistic, generic presumption (Natural meaning argument): "Implicit in the terms "household member" or "guest" is that access to the premises has been granted by the tenant".
4. Structural presumption (Historical argument): "In Memphis Housing Authority v. Thompson, the Tennessee Supreme Court noted that the statute, and the lease provisions that were derived from the statute, refer to four separate categories of people: (1) the resident; (2) household members; and (3) guests or (4) other persons under the resident's control".
5. Pragmatic presumption (psychological argument): "In an effort to end what Congress termed the "reign of terror" imposed by drug dealers on public housing tenants, 42 U.S.C. 11901 (1994 and Supp. IV 1998), Congress enacted Section 1437d (l) (6) in 1988 and strengthened it in 1990 and 1996. A different interpretation deprives public housing authorities of an important tool to achieve safe and livable public housing, and to deprive public housing tenants of protection that Congress found to be of central importance for their security and well-being".

5. Legal and Linguistic Knowledge as Arguments in a Legal Case

As can be seen, what is intentionally called a legitimate "contextual" genre knowledge is an array of linguistic, pragmatic and legal knowledge far vaster than micro-linguistic surface features and linguistic markers typical of statute or legal language. As befits a cognitive interpretation of legal communication and interpretation, we must stress the genre-induced rules of interpretation as a genre-determining type of intended cognitive environment. Genre is not determined by specific selections of surface language materials.

I offer two illustrations of how global interpretive principles operate here, to demonstrate how interpretive strategies are "switched" on (long) before we approach an utterance. Consider, as our first example, prose fiction, with its narrated details. In this case, there is an underlying assumption that the factors selecting, from all the globally possible details, only a tiny fraction as worthy to be presented as stimulus for cognitive belaboring are

intentionally selected by the author and will be instrumental in the progress of the story (Chafe 1992, especially "fine-grained resolution" pp. 237–39). This presumption is one of the uppermost determinants of comprehension and interpretation of prose fiction.

A second example can be drawn from Skoczen's (2021) view of the implicatures in statutes. Skoczen (2021) argues that Grice's cooperative principles are insufficient to make explicit what happens in legal interpretation. She postulates an additional "fixed strategic normative framework" (Skoczen 2021, p. 4), including a dominant "strategic maxim", superimposed on the still functional Gricean maxims, as determinative of legal interpretation. So, the domesticating powers of genre do include micro-linguistic means, but they can act only in the strong company of all other kinds of more abstract genre determinants, ranging from the "strategic" maxims to the notion of "intention". Skoczen's (2021) own example is from a case of harassment, where she argues these principles are operative alongside the problem of uptake in this genre, as well as the register factors determining options in choice of expressions.

After all, there is already consensus that the topmost determinant of genre is "intention" or "purpose", be it in the minds of one person or of several people—a mens rea—or a more abstract societal purpose. After all, there can be no purpose unless it is intended by individuals, whether concretely or abstractly. The individual may not even be consciously aware of having a purpose. Individuals may, for example, gang up and engage in mobbing actions, resulting in "mobbing" as an incipient genre type, defined by the purpose of jointly denigration. Or "society" (a composite of minds of individuals) may evolve such a purpose. Over time, society might package a purpose-defined configuration of language-system resources, pragmatic comprehension and interpretation conventions into a complexity-reducing, automated communication package that defines genre-relative normality expectations, such as those reflected in Horn's I and Q baseline principles that are part of Kahneman's System 1.

The dualism between the linguistic surface and the non-surface pragmatic strategies is explicitly recognized by Palermo and Visconti (2023, p. 250), who argue that genres "are governed by diachronically stable and generally universal principles, as they relate to general cognitive and communicative features". For Palermo and Visconti, these features manifest in the shape of cohesion and coherence-based elements of genre definition, while the approach in the present paper gives clear logical and functional precedence to the coherence, that is to say, to the pragmatic side. My privileging of the pragmatic side seems also warranted by the inclusion of a historical perspective on the phenomenon of genre. From history, we can see that genres come and go; sometimes their purposes become opaque in retrospect. On the other side, genres may be in statu nascendi and on the way to "sedimentation" (Palermo and Visconti 2023, p. 249), a process that welds the linguistic and pragmatic elements into an unanalyzed whole to function as a composite element of System 1.

The historicity of genre formation is strikingly demonstrated by Rosenberg, through the application of the relativity of genre formation ("advertising-as-genre") with respect to historically arising purposes on an abstract societal level. In an analysis of the history of advertising from the 19th century, Rosenberg writes the following:

> As a genre, made of texts and images that dialectically forge and are forged by changing media and audiences, advertising is not an obvious object but a historical problem and ongoing effort of creation, in which law participates.

From the information-theoretic point of view, what is essential is that it is only the whole package that is called into service, not discrete elements. Only then is the package capable of solving the informational bottleneck problem, from both the production and the comprehension side in communication. The individual modularized element contributes little towards informational capacity and efficiency. Nor do isolated elements have explanatory power in a modular analytical investigation, except when seen through their relative function within the larger genre whole.

6. Making Meaning: Bottom Up

Having emphasized, in the preceding section, the top-down character and effect of contextual knowledge that is "switched on" by entering the communicative game of a specific genre, it is now time to include the bottom-up perspective in construing meaning. After all, the default mechanism of comprehending involves the sequential processing of sentences as they are transformed from physical inscriptions to sequences of morphemes and constituents. Modern pragmatics has conceptualized the way meaning is built up from, and triggered by, incoming sentences in terms of staggered processes of "meaning making". How this process is conceptualized is of particular relevance for the societal domain of law, which has the construction of meaning on the basis of texts as a constitutive linguistic basis of its activity. For this reason, some aspects of the following discussion pay special attention to issues of legal interpretation, especially those relevant to genre-based top-down processes. The discussion will touch on a number of theory-internal issues, for which we can point to a rich body of literature. The interested reader is invited to consult the survey of pragmatic linguistic theory offered by Jaszczolt (2023).

The initial template of our discussion will assume, at the beginning of meaning extraction, the availability of a "text sentence", or a compositionally available raw proposition, composed of language-internal morphemic, lexical information chunks, in the sense of Lyons' (1981, p. 196) "system sentence", but which is made more explicit by the addition of anaphorically given information. There would then be a first level of meaning extraction without explicatures. Not even Recanati's primary enrichments as they are triggered by the componentially available "text sentence" would be needed.

> Saturation is a primary pragmatic process. If the uttered sentence is 'She is smaller than John's sister', then in order to work out what is said, I must (at least) determine to whom the speaker refers by the pronoun 'she' and what the relevant relation is between John and the mentioned sister. Were saturation a secondary pragmatic process, I would have to proceed in reverse order, i.e., to identify what is said in order to determine those things. Beside saturation, which is linguistically mandated (bottom-up), there are, I claim, other primary pragmatic processes that are optional and context-driven (top-down). The paradigm case is free enrichment, illustrated by example (1):
>
> (1) Mary took out her key and opened the door.
>
> In virtue of a 'bridging inference', we naturally understand the second conjunct as meaning that Mary opened the door with the key mentioned in the first conjunct; yet this event is not explicitly articulated in the sentence. Insofar as the bridging inference affects the intuitive truth conditions of the utterance, it does so as a result of free enrichment. In typical cases, free enrichment consists of making the interpretation of some expression in the sentence contextually more specific. (Recanati 2003, p. 23f)

Primary processes like saturation (linguistically mandated) and free enrichment (pragmatically mandated) are essentially automatic processes, operating below conscious awareness. They are, therefore, arguably part of the speaker's System 1, in Kahneman's terminology. Primary processes are not defeasible. These first analytic, logical steps in meaning making result in a layer of meaning that is falsifiable, variously referred to as the proposition "what is said" and "explicature".

Such primary stages of meaning extraction already require access to pragmatic, language-external knowledge. Only the initial construct of the "system-sentence" is, by definition, a completely a-contextual and a-pragmatic, purely logical-semantic level of analysis. The recognition of this initial level of knowledge identification would, as a logically conceived process, function as the first step in triggering the next steps. After all, if there is not a linguistic form awaiting saturation, the later processes cannot logically happen at all.

What is important, at this point, is the inclusion of selected elements of pragmatic external knowledge from the very first stages of the meaning-building process in the shape of the first inferential processes.

The proposition—in the above-described sense—is then what triggers, in an extremely context-dependent way, further implicatures of the kind variously called "particularized" or "conversational". Their hallmark is that they are in this sense "postpropositional", that they are defeasible, and that they arise on the basis of the proposition or the explicature.

To give an example, taken from Carston and Hall (2012, p. 47f):

(1) Max: How was the party? Did it go well?
(2) Amy: There wasn't enough drink and everyone left early.

Amy's answer as in (2) will be freely (secondarily) enriched to yield the following explicature (3), which we will at this point accept as the "proposition" or "what is said":

(3) THERE WASN'T ENOUGH ALCOHOLIC DRINK TO SATISFY THE PEOPLE AT THE PARTY AND SO EVERYONE WHO CAME TO THE PARTY LEFT IT EARLY.

It is only through the explicature as in (3) that the ultimate intended meaning will be arrived at:

(4) THE PARTY DID NOT GO WELL.

Amy's response triggers, as a post-propositional process, the conversational implicature in (4). (4) is the ultimate pragmatic meaning of Amy's answer, implicated to be so construed by the speaker, and only vestigially related to the content of the proposition, or even to the system sentence.

By way of another example, taken from the rich literature on the subject, consider the following utterance in the context of the meeting of a selection committee:

Text sentence: "She has a brain".

1. SHE HAS A HIGHLY FUNCTIONING BRAIN (explicature and proposition, what is said)
2. SHE IS THE IDEAL CANDIDATE FOR THIS POSITION (implicature)

In both examples, the (correct) implicature is the purpose of the utterance. It must therefore be included in any realistic explanatory account of communication. The totality of these meanings goes, since Grice, by the name of "meaning nn".

Meaning construction departs from this level of "What is said", finally going on to a level of "Meaning nn" that subsumes all of the meaning effects (presumed to be) intended by speaker and (legitimately) construed by the hearer. From an early stage, it develops towards a final public putting-on-record that the applicant in question is the ideal candidate for the job.

What should be clear by now is that there is a long stretch of inferential distance from the code input (system sentence) to the final implicature. The length of the distance presents a challenge to any account of meaning extraction in a domain like law, which operates on its self-confidence in being able to make explicit how utterance meanings (meaning nn) are arrived at, along a stretch of linguistic communication from forensic linguistics via police work to subsumption and legal interpretation.

It should be said that the exposition given in the above section is embedded in the history of the "border war" issue between semantics and pragmatics. As it happens, ever more information in the meaning-making process was "taken over" by pragmatics, with an endpoint in the shape of "contextualism" (Recanati 2003). For contextualism, a maximum amount of information comes from text-external pragmatic knowledge, so that the emphasis is on pragmatics. To stress "emphasis" does not imply that semantics are overlooked, nor to belittle the contribution of semantics to interpretation, but to relativize it—and yet nonetheless to emphasize that semantic understanding is arrived at by pragmatic means. It is likely that none of the statements in the foregoing would be fully subscribed to by any of the theory protagonists named, which is usually the case when applying categories of theory to an applied field such as law.

Apart from the theory-internal reasons to make pragmatics the first point of departure in making explicit the construal of meanings on all levels, it would appear that there

is significant added value in applying pragmatic categories of analysis to the practical conduct of law. It is very much to the credit of the ground-breaking work of Smolka and Pirker (2016, 2018) to have shown how knowledge of the issues in this field can inform the processes of legal interpretation and help shed unhelpful ideas and ideologies of language still rampant in the legal professions.

7. What Is in the Text?

A standard issue in legal theory is the question of to what degree the meaning of legal text resides or should reside "in the text", and to what extent meaning involves accessing contextual knowledge. The branches of this discussion go under a variety of names, "textualism", "intentionalism", "originalism", and others. It is natural that a theory of meaning origination that has at its center the relative shares of knowledge from language surface versus from pragmatic context should make this issue a subject of inquiry. It is important, for reasons of its own efficacy, that the field of law defines precise points in the meaning origination process at which extralinguistic knowledge can, or even has to be, legitimately accessed. For pragmatics, it is axiomatic that not all meaning comes from "the text" and certainly not from the level of the "system sentence" where linguistic meaning is self-sufficient, pure, and pragmatically undefiled.

The phenomenon of underdetermination has been amply discussed with respect to how it operates at the "bottom level", i.e., where specific linguistic and syntactic structure undergo multiple interpretability and give rise to persistent types of vagueness (Witczak-Plisiecka 2009).

The notion of "underdeterminacy" has something of an internal history in modern Gricean, neo-Gricean, and post-Gricean theoretical discourse (Jaszczolt 2023), which does not concern us here. We will simply accept that there is no level or representation or meaning derivation that is not affected by pragmatic knowledge. Pragmatic knowledge is always present in non-linguistic knowledge mutually assumed to be present, be it through indexicalization, reference, or modulation. The exact contribution from each domain—or the location of the borderline between semantics and pragmatics—does not concern us here, beyond the statement that knowledge from both sources is available at every point, including knowledge accrued during the processing of a text.

Linguistic indeterminacy is a classic issue for language and the law. For the purposes of law, indeterminacy is always seen as pathological, a recurrent point of crisis for legal professionals. This conception—and practical experience of—indeterminacy is part and parcel of language ideology in law. Of course, lawyers' linguistic heaven would be where the "text" is fully self-sufficient, its meaning a fully autonomous effect of surface forms. No grueling interpretive efforts would be required. A lawyer's heaven would be imagined as the following:

> a "paradise "where all words have a fixed, precisely ascertained meaning; where men may express their purposes, not only with accuracy, but with fullness; and where, if the writer has been careful, a lawyer, having a document referred to him, may sit in his chair, inspect the text, and answer all questions without raising his eyes. (James Bradley Thayer, cited in Tiersma (2010, p. 29))

And the law would be as follows:

> a code at once so flexible and so minute, as to create for every conceivable situation the just and fitting rule. (Solan and Tiersma 2012, p. 88)

And this would be computational linguists' heaven, too. On the lowest level, there is a "language", with its inherently indeterminate features, such as—inherent dangling participles, scope ambiguities, and the like. Such cases are amply discussed in the linguistically oriented literature on legal interpretation, where an entire case may hinge on what is called the "ambiguity" of, say, the connectives "and" and "or". The fact of inherent indeterminacy has long been recognized, and the issue is a key part of any competent introduction to the linguistics of statutes.

If we want to interpret a statute—or any text—the first question is the following: what is it that we "have"? The very notion that we "have" something in the surface linguistic forms and features of a text is already problematic enough. The title of a famous article on literary interpretation reads as follows: "Words are all we have" (Schaub 2010, p. 185). To law, pragmatics must retort: What exactly do we "have"?

What we "have", once we know we are engaging with a piece of discourse in the legal domain—let us say of a statute—is all the knowledge that resides in the genre "statute" and the super-genre "law", such as described in the above citation by Giltrow. We also have higher order knowledge, such as referred to by Rosenberg. Essentially, what we have are the genre concepts and various kinds of interpretive principles, be they obligatory or be they moot.

It may be seen as definitional for types of legal genres that there is an asymmetry not only in types in discourse rights (like topic determination, turn-taking, etc.), but in terms of cognitive environments, as e.g., in depositions. Here, the presumption is that the deposition takers do not have the same knowledge as the deposed persons, whereas in a cross-examination, the cognitive environments likely coincide. Questions asked in cross-examination are not intended to add to the cognitive environments of known facts.

Apart from the issue of the "richness" of the external knowledge, another issue seems to be that these constitute prepackaged knowledge types. An occurrence of a concrete text or discourse is always an instantiation of a type. I would even go so far as to say that you cannot make an utterance that relies only on internal knowledge. Whenever you speak, you invariably speak within a genre.

So there is, to start with, something like a "logical" primacy of genre. What is there first is really not a sentence or utterance, but a genre. Genre is what we "have" first, even before we read the first sentence or hear the first utterance.

In addition, we have to assume not only a logical primacy, but also a psycholinguistically and processing primacy. Genres these pre-existing knowledge complexes, or packages, are available to the hearer, and assumed to be so by the speaker, before any other decoding takes place. Before you read or listen to a text, you know what genre you are engaging with. It is like turning on a specific light, in which you see the incoming figures. The "words" never walk alone.

The language-using organism is not a lonely organism; she is in a lot of company and cannot help being so. Our modeling of what happens in construing meaning should reflect the psycholinguistic primacy of the genre. Our perspective here, as proposed by Jaszczolt (2023), supersedes a more traditional view of meaning making, namely the so-called "pipeline" view. The pipeline metaphor aptly characterizes meaning making as a logically ordered sequentiality, where one step of meaning construction depends on the availability of the former respective "lower" level. In the pipeline view, secondary processes like implicatures are only possible after the existence of propositional explicatures, and the latter themselves only after free enrichment processes have taken place.

Seen from the point of view of the expectations, as current in law, that the meaning is "in the text", the obvious question from the pipeline point of view is the following: exactly where in the derivational stages is "in the text" located? Each derivational stage offers a different answer, with some answers more or less attractive from the perspective of law (cf. below).

For those who accept this logical sequentiality, with information required at any later stage available earlier stages, the pipeline forms an elegant hypothesis for a real-time, orderly process. If the processing perspective itself forces a concept of genre as a psycholinguistically and informationally plausible concept, it comes at the cost of quantity–arguability. This type of "cost", in the sense meant by Giltrow, is a cost against the background of what I have termed a language ideology that sees language predominantly as surface forms. My intention with the term "ideology" is not pejorative; I appreciate that many applied pursuits, such as computational approaches and translation, must see language this way, of necessity. For pragmatics, however, such ideology entails the

unscientific default expectation that all manner of content and function needs surface signaling, which is incompatible with the arguments about informational bottlenecks advanced by Levinson (cf. above § 2). The absence of redundancy in the shape of non-linguistic knowledge, with all information signaled through code, would in fact make communication impossibly unwieldy. Arguably, 80% of what functions in communication is context, and it requires our package deals to function at all. Hence, the primacy of genre, understood as prepackaged, subconscious knowledge types.

A lot of external knowledge is adduced at several or all stages, depending on the nature of the genre. Where does it all come from? Does it come from a preselected range of genre knowledge? The major determinant of selection is the relevant genre, in the broad way we define it here. It is indeed accessed and selected in an orderly sequence, such that if one type of knowledge is called in at one stage, then it will then require another, or next, element of knowledge. Other elements are present the very moment genre is "switched on", mostly through external, situational socio-cultural knowledge. The external triggers are contextualization cues which will affect the "modulation of lexicon, grammar, and prosody, the very contextual frame within which it should be interpreted, a bit like a snail that carries its own house around with it" (Levinson 2024, p. 34, referring to Gumperz' notion of such cues).

One logical question to ask is the following: At each stage, as a different type of knowledge is accessed, is the prior stage discarded? Is each type of knowledge, at each sequential stage, accessed and then discarded? Is the process akin to sequentially reaching into different drawers of a cupboard, each in turn? With so many stages involved, if meaning making really operated in this way, then it would, again, be less efficient than we already know to be. It would take too much processing time and exacerbate informational bottlenecks. Processing constraints therefore force a view that different "drawers" are made salient at different points in the process, and the end result of a "discourse", the final utterance "point", is not computable by individual steps. Instead, it must have an element of supersummativity and may even be renegotiated interactively if the genre allows, as it does in conversations. For the time being, this supersummativity will have to remain an element of "unknown dark matter".

Elder and Jaszczolt (2024, p. 9f) represent an advance on earlier, more or less strictly "pipeline" views, such as are still basic in Relevance Theory. In their "Default Semantics", Elder and Jasczolt explicitly acknowledge a wide range of types of external sources of information that may be accessed selectively. Their approach ultimately results in a "flexible functional proposition":

> The merger of information coming from different sources, through the associated processes, then produces what are called in Default Semantics *merger representation* [. . ..], which allows us to represent the composition of meaning that is arrived at by the interactants themselves.

Importantly, what is included in the range of "external" information are not only static types of knowledge, but also processes, including processes of interpretation and modification of knowledge.

All types of enrichments, modulations, primary and secondary, as well as all types of implicatures, require access at so many points in the comprehension process (and are calculated by the speaker to be accessed by hearer) that it suggests a top-down genre perspective. Without genre, as Levinson asks: "How does a recipient find from the forest of possibilities just the implicatures intended within just a few hundred milliseconds?" (Levinson 2024, p. 22). Levinson (2024) discusses the processing and information-theoretical perspectives primarily from the point of view of spoken conversation. But all the issues Levinson discusses apply in other kinds of language use, given some modifications for the case of written discourse and for the handling of particularized implicatures, both of which may involve some amount of conscious System 2 "thinking". Hard legal interpretative issues of course require conscious reflection, typical both for legal statutes and literary study.

A processing perspective leads inevitably to a genre perspective: the relevant knowledge types are strongly constrained and restricted in such a way that they can plausibly and manageably be accessed at any time with minimal processing effort. So the pre-packaged knowledge is present right from the start of processing. Here is where a classical, explicit "pipeline" view of knowledge processing, while logically elegant, is implausible from a *psycholinguistic* point of view, for reasons of strict capacity limits and speed of processing (cf. also the argumentation in Kuhlen and Abdel Rahman).

One ramification of this psycholinguistic processing argument is the "functional pragmatic proposition", suggested by Jaszczolt (2021, 2023):

> So, we need to go a step further. We need a unit which belongs to speakers and addressees and reflects their conversational interaction in 'meaning-making'. But; moreover, it also needs to pertain to truth-conditional content as it is understood in contextualist truth-conditional theories of meaning. For this we move to the concept of a *functional proposition*. (Jaszczolt 2021)

A functional proposition is a structured proposition that reflects the composition of the main communicated meaning. It captures the primary intended, recovered, and partly co-constructed meaning as it is understood by interlocutors. As such, to reiterate, it captures the primary communicative *function* of the utterance. This primary meaning can be directly or indirectly communicated and may or may not correspond to the speaker's initial intended meaning, which is negotiated interactively. In addition, the structure of the proposition relies on the varied, multimodal informational input in communication. That is, it relies on information about meaning that comes from different sources in communication, not only from the utterance itself. Rather, we communicate by immersing our utterances in a situation that exploits socio-cultural defaults, background information (i.e., common ground) and other information sources, as will be discussed in more detail in Section 8.

A functional proposition contains all interpretations, forces us to give up the "pipeline" as psycholinguistically unrealistic, implausibly cumbersome. So down the drain goes, in psycholinguistic terms, what were once thought to be the autonomous material of "what is said". Instead, "what is said" is now identical with the functional proposition. It is the implicature "She is the most suitable candidate" that is actually "what is said", an idea not easily entertained in the legal profession.

Even with the functional proposition understood as "what is said", it is not the case that the sentence or utterance is prior, and that the pragmatic aspects of meaning, i.e., genre knowledge, are then "added", incrementally called upon, operated on either in a logical model sequence or a real-time sequence. What is there first is genre and all knowledge, in real, infinitesimal time.

8. Meaning Making in the Law

For theorists, at least, the pipeline view of meaning making has a strong competitor. Yet, the classic concept of sequenced meaning building remains stimulating for linguistics and legal linguistics. "Literal meaning" is, of course, a perennial issue in law. Conceptually, the "literal" could be located after disambiguation: narrowing, modulation, and indexicalization, at the stage of the text sentence or the proposition located at the level of the explicature. If so, it begs the question: Are *scope ambiguities* resolved by a principle, on the spot? Are they resolved by a canon of interpretation rule, as part of the genre package? Or by larger considerations of plausibility of the resulting solution, i.e., made by a process further along the sequence?

In law, the issue of literal meaning is especially relevant for expressions like "and" and "or" (Smolka and Pirker 2018). Do we assume one literal meaning—as Smolka and Pirker do—and let pragmatic principles—in effect inferences—decide at *a later* stage in the derivation (as explicatures?) which meanings to apply, or do we assume a proliferation of separate langue meanings (system sentence) to be available at different points along the derivational path?

It turns out that the pipeline model retains some interest for the teaching of the subject, such as the importance of "stages" in this process. Different legal interpretations of language indeterminacy can be conceptualized as being resolved in a staged process. In this framework, the question becomes the following: Where is the proposition located? What part of the proposition carries a truth value? Is such a question needed for linguistic inquiry in legal procedures? A related issue is how long earlier derivational stages are still "available" in comprehension (Recanati 2003 and can be made explicit as a criterion in establishing "what happened" or "what was said" in language crimes (such as hate speech, defamation, incitement, and the like). Of course, the applicable version of "what is said" or the "proposition" would then be the "early" one as the result of primary pragmatic processes, as the explicature, or even text sentence, and not the functional proposition. Recanati (2003, p. 29) explicitly points to the possibility that "derivationally earlier" stages are in fact "skipped", and given the fully available contextual information, a directly functional proposition in Jaszczolt's sense is interpreted:

> An important difference between the Gricean model (according to which the literal interpretation is processed first) and the parallel model just outlined is this: on the parallel model it is possible for an utterance to receive a non-literal interpretation *without the literal interpretation of that utterance being ever computed*. The non-literal interpretation of the global sentence does not presuppose its literal interpretation, contrary to what happens at the constituent level. If the non-literal interpretation of some constituent fits the context especially well it may be retained (and the other interpretations suppressed) *before* the literal interpretation of the sentence has been computed.

So the "availability" (Recanati 2003, p. 20) of earlier derivational stages is not in every case to be presupposed. In fact, it may have to be retro-construed as a task of Kahneman's System 2.

"What is said", in the older sense of a derivationally earlier proposition, can matter a lot in courts, especially with lawyers being so preoccupied with the "logical content" of what is said, ascribable to words as such, as something "objective and therefore, it would seem to all involved, demonstrable. The prosecution or defense in cases of language crime, such as hate speech trials, runs on shallow ground when trying to base a legal strategy on an older notion of "what was said" (Guillén-Nieto et al. 2023).

The specific language-ideological view that permeates thinking in the legal profession is aptly characterized by Smolka and Pirker:

> While international law literature puts much emphasis on interpreting a "legal text [. . .] in such a way that a reason and a meaning can be attributed to every word in the text" (Linderfalk 2007, p. 108, quoting Haraszti, emphasis—the notion of word appearing to be legalese for conceptual term (Linderfalk 2007, p. 106))—the notion of text seems not worthy of any definition altogether. This may have to do with the fact that a text is not "available" or "readable" independent of interpretation, in the process of which one may then be busy focusing on utterances. The question is as follows: given that a text typically consists of a sequence of utterances, how does this affect the interpretation process? (Smolka and Pirker 2016, p. 30)

This, then, is the attempt to construe meanings exclusively bottom-up, as follows logically from an assumption of a full-as-possible autonomy, as is standardly assumed to apply in statutes of other legal discourse under the presumption of literalness.

The effect of genre is naturally also felt in the constraints it places on the meaning of non-propositional expressions. For instance, the semantics of "whereas" in setting of paragraphs, or the meaning of the conditional "if" in statutory regulations (Szczyrbak). Conditionals, theticals, discourse markers, and other elements, have characteristic meanings that are special to particular genres.

For instance, tense change is of particular interest in forensic linguistics. In forensic linguistics, tense change is often taken as indicative of lying, or that the statement is simply not true. But in specific types of narrative, those same markers can have the opposite meaning for truth evaluation. Instead, it is the *absence* of tense change, and the expression of emotion at a narrative peak—an eminent structural and highly diagnostic point in a narrative—that is indicative of lying.

The emphasis here is on what *can* have: genre provides the "explanation" of why we construe this meaning rather than another. It is built on the knowledge present at whatever point you are at in the particular genre in which you are involved.

This narrowing and specialization of meanings is just another example that raises the question: "How can we know so little, given that we know so much?" The plethora of meanings—the "forest of possibilities"—of words like "and", and of grammatical forms like tense changes, is reduced by the local type of interaction.

One would expect that move-structure marking, while being the rhetorical core of suasive legal genres, will not be marked. How do hearers know? They are a priori willing to interpret next sentences or utterances as next moves, depending on the connectionist status of the sentence, as predicted by the genre. So the expectation of segmentalization, of surface marking as a default, is misguided on the lower level. To the extent that the corpus is computer-based, there is evidence that, due to the higher redundancy through the technicality of the affordances, there is even less surface marking generally in Internet language.

An interactionist view, where meanings are also top-down co-determined, is communicationally adequate. What you interact with principally is genre knowledge. Genre knowledge includes, of course, terminologies, including legal terminologies with all kinds of vagueness, like in legal standards.

We must assume that the priority of this knowledge in logical and processing terms also applies to, and is indeed a criticism of, the concept of a macrostructure, as it is manifested in the text in the Appendix A. The implication of the formation of the macrostructure is that it is linearly and sequentially constructed on the basis of the incoming propositions. While this may be an elegant concept to explicate propositional meaning, it is, as I have said, psycholinguistically and pragmatically unrealistic. On an incrementalist step-by-step assumption of m-structure building, it is clear that it would simply take far too much time. However, as a logical structure for thinking or teaching about meaning, it is very useful indeed. The text in the appendix possesses both a macro-structure and super-structure, as part of the text's genre character: opinion text, rhetorical expressive/directive speech act.

Since functional structure is a hallmark of many genres (narrative, legal, etc.) and part of the "switched on" cognitive environment, we cannot be surprised that explicit structure boundary marking is also redundant. Genre knowledge makes us perceive the sentences at the beginnings of the four paragraphs as topic sentences, enabling an immediate expectation that what follows will be a fleshing out of the topic sentences. The reader knows many things before even the first morpheme is decoded. Contributing to this is a meta-relationship of a causal nature: the reason why I am saying/claiming this, hoping to convince is that ("topical sentence"). The reason for my assumption is then the rest of the paragraph, which is processed as giving reasons for the thesis put forward in the topic sentence. This discourse structural fact plays out on the level of relationship between sentences: sentence connection does not obtain between first sentences and what follows.

The text has no surface structure marking, as it is redundant for the reasons just set out. The writer here can well afford to *not* mark this structural meaning on the surface.

Given the fact that a lot of "pre-propositional" pragmatic knowledge is part of what is said, and that there is an ongoing controversy about how up or down the "what is said" is to be located, it seems appropriate to include all manner of knowledge functioning in a new concept of what a "proposition" is. It is necessarily one that reflects all types of meaning, and which is "functional" in a bounded piece of discourse, and that includes, but not exhausts, so-called intended speaker-meaning.

Although obsolete from a scientific point of view, it may be useful to keep the notion of "literal meaning" as an "operational" concept for applied purposes, just like all other similar notions, like "basic meaning", the notion of an autonomous "text", and so on. But, as I have argued, it will be interesting to see how this can be profitably exploited for explicating and elucidating the handling of meaning in a legal context.

9. Logical vs. Real-Time Ordering

As has been adumbrated here, the logical ordering in the "pipeline" view is, from a real-time processing point, entirely implausible. What is actually there first, in real time, is the situation and the communicative intention, present long before any proposition or system sentence is on the table.

Even earlier ideas about meaning making, assuming that text processing and comprehension started with availability, first, of code information, also assumed the presence at the prior point of non-linguistic knowledge that would enable processing of first stages in the first place.

What is there first is the genre, not the language: this view of the primacy of genre is compatible with the idea of a "functional" proposition in the sense meant by Jaszczolt. It is also compatible with Recanati's (2003) view, articulated in the following:

> As I have been at pains to emphasize, the meaning of the whole is *not* constructed in a purely bottom-up manner from the meanings of the parts. The meaning of the whole is influenced by top-down, pragmatic factors, and through the meaning of the whole the meanings of the parts are also affected. So we need a more 'interactionist' or even 'Gestaltist' approach to compositionality. (Recanati 2003, p. 132)

Genre, with its superstructural internal redundancies, will so strictly constrain any possible next move in a sequentially ordered type of discourse that it is unnecessary to mark it on the surface. Indeed, offering additional signaling materiality would go against expectations in such a constrained local environment. It is part of the interpretive schema in this genre that such semantically conditional structures will be interpreted as conditional in this genre. It would be a waste of signaling materiality to specially mark them on the surface and potentially would be a violation of a communicational maxim of quantity.

This semiotic-economic interpretation accords with a more general view of the relationship between surface and cognition/interpretation in genres. Following the discussion of *underdetermination* above, it is a well-known fact that certain genres are only minimally defined by linguistic surface information, a phenomenon we designated as "genre-maximalism" above and with Giltrow's dictum that "form alone" (meaning surface, code marking) does not define genre.

Levinson (2024, p. 30) refers to a more general principle formulated by Sacks and Schegloff: "oversuppose and undertell". While originally with reference to spoken conversation, the principle can be taken as operant in managing any knowledge with the help of genres. Sacks and Schegloff's principle implies an economical procedure, achieved by "amplifying coded content by virtue of prearranged rules of thumb or pre-packaging of default assumptions" (Levinson 2024, p. 21)—rules and assumptions such as constitute genres. What this amounts to is a major relativization—and, comparatively, an acute downgrading—of the role of information supplied by code, i.e., by language. In other work, the process has been referred to as the "de-surfacing" (Nicklaus and Stein 2022, pp. 157–77) of analytic concepts, especially as far as the discourse level is concerned. It is a misguided expectation that functional structure parts, as in genres with a pronounced superstructure or conversational moves, would imply that a central information management technique as a design feature of language could not work in the efficient way it does. Still, such beliefs are widespread elements of ideologies of language, such those ideologies still dominant in law.

In some genres "undertelling" seems to apply with a vengeance, especially in artistic, or literary genres, where a minimal definition appears to have taken maximal recourse

to non-surface definitional parameters. Just as contracts, for example, are defined by speech-act pragmatic categories, the same is true for literature. Literature is not defined by surface features, but by a specific type of "cooperation" with what is "in the text", however difficult this has been to conceptualize (cf. above). The location of this meaning level (the proposition) may itself be specifically definitional for literature (one is tempted to say "genre" here). Literature may in fact be a special case for extreme minimalism, as it cannot be surface-linguistically defined even in the most limited way.

Further support for this assumption might be derived from the fact that literature can be generated by machines like ChatGPT. It is not possible to define surface constraints for verbal material that would be definitional for what is machine-generated or what is human-made. It is—as is normal for modern artistic production—the interpretive act that defines what is art. Even the nature of the type of inferential processes (like their non-finiteness in principles, the non-availability of an end state, as in principle underlying relevance theory) is open and non-determinate.

A similar perspective applies to the discussion whether ChatGPT in fact generates *new genres*. The question is an extrapolation of the issue whether mediality is definitional for genre boundaries. To some extent, mediality surely is definitional. The blog is a child of the Internet (Puschmann 2010), with parentage in the form of the written diary. Arguably, the law, as an abstract canonical body, does not change its nature and its canonical content only by "packaged" in different new media (Greineder and Stein 2023). Interesting questions are nonetheless being raised for a number of legal genres relating to new genres or genre identity in a new medial garb (Anesa and Engberg 2023).

To the extent that the notion of "intention" is part of a genre characteristic, the issue is slightly different. The question is, instead, the following: can the recovery of the producer intention be at the center of meaning making in ChatGPT-produced text? Given that the producer intention (in legal theory, "intentionalism") is an important dimension in analyzing texts in law, what does this imply for machine-generated text, given that our notion of intention is that it is something only a human can have? What does it imply for different types of genres in the legal domain? What about statutes? And what about "criminal" texts, such as those under suspicion as "hate speech" (Guillén-Nieto et al. 2023), where a *mens rea* is crucial for establishing a "fact" and actionability? Hate speech itself is arguably not a genre, although it is strongly tied to intention, and the "linguistic" (surface, code) share of meaning determination can be less than minimal in some cases. By way of a speculation, does "declaring it" (linguistically analyzing it as) a separate genre, tied to contextually documentable intentions, make it easier to hold humans who have used a machine accountable? If so, forensic linguistics would then face a very different type of task.

After the foregoing discussion, the very notion of "genre" naturally raises some more fundamental issues. As I have said, genre, as we understand it, is a concept from cognitively oriented pragmatics. If "text type" considered pragmatic knowledge as a patching-up repository of "context", genre theory in turn considers surface information as only residually essential, stronger idiomatically, and idiosyncratically determined. Surface information is computed not as types, but only as tokens.

To summarize, there are several related notions of "genre", with (in the order given) descending degrees of "surface-relatedness":

1. A macro-sense (Rosenbergian) in the sense of law or science being genres.
2. A slightly lower level that ties genre more specifically to activity types in specific social or institutional situations (in the sense of Engberg, or activity types as Levinson 1992).
3. A mid-level sense (as implicitly advocated here), which would agglomeratively be representative for the macro-level.
4. A low-level notion, closer to individual surface form-oriented notions, that often deals with didactic considerations (how to teach genre and or about genre) and which is in a way pre-pragmatic, like the early work on Genre by Bhatia (2023) and at which more surface-dependent notions, such as computational work, must operate. Dorgeloh and

Wanner (2022), for example, contribute a valid, corpus-based analysis of surface forms together with their genre privileges of occurrence.

Generally, surface forms are epiphenomena of pragmatic parameters, types of interaction and embeddings in situational, actional contexts. Illocution types would seem to be very important, especially for genres that make up the world of the law in the Rosenbergian sense.

A basic determining constellation for a genre type 2 might well be a type of situation (as part of a no 1 dimension) with its own types of illocution and specific interactive and asymmetric discourse parameters, like a police interview. This function in turn determines the type of "possible content" or "possible move". What can come after what, and with which function, would be quite predictable and would require little if any surface marking. Levinson (2024, p. 18) points out that sequentiality is a key indicator: "A central observation in CA is that sequential context in a series of actions is often a powerful heuristic".

Another direction of radical questioning would ask if it is possible to have genre without any verbal share at all. In contrast to the so-called Sorites paradox (how many grains do you need to minimally have to form a heap? Answer: just one) it would have to be argued that, given the generally subordinate function of surface materials, the limiting case would indeed have to have genre without any verbal share whatsoever, which of course does exist and not only in the field of arts.

If this kind of argumentative radicalization of the theoretical issues is useful for heuristic argumentation, like in any other conceptual context, it yet does not damage the practical ineluctability and the theoretical soundness of the concept of genre. After all, you can also ask the following: "Can you have a phone without representing a phoneme?"

The pervasive share of gradience raises the issue of the "operationalizability" of the concept. The point can and has been raised that what is a genre cannot be "strictly" so, with categorical yes–no boundaries. Neither can a genre be algorithmically defined, as appears to be a precondition for the "new discreteness" postulate of much modern science. Genre cannot be defined under the dictatorship of quantificational approaches, those giving priority to discrete boundary scientific categories, no matter what sense it might make in any individual case.

Many, if not all, of our linguistic categories are fuzzy and gradient, from the phoneme all the way to concept of an NP ("nouniness"). So this cannot be used as a "principled" objection against a gradient concept of genre.

It is perfectly true that the operationalizability issue may be an insurmountable difficulty, except for genres that do have a high degree of surface representation. It would appear that genres in the domain of law, because of their high degree of ritualization indeed do not offer themselves readily for analysis in terms of quantifiable surface markers. Ritualization implies contextualization, and this in turn that less surface signaling is necessary. The "lower", and well-embedded the units are, such as conversational moves, the higher the level or redundancy of what is possible and expected. Another feasible and common procedure is to work with something like a first derivative or watered-down version of the genre concept and to posit a type of phenomenon closer to a text-type as genre-akin (close to no 4 above), making it amenable to automatic analysis; however, there are some caveats as to a functional interpretability of statistical results and preferences in terms of genre.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study.

Conflicts of Interest: The author declares no conflicts of interest.

Appendix A

Newspaper Opinion Text

- A Wilsonian Fog over Import Policy
- 1 In his ramble around the economic situation last week, the Prime Minister again broached the subject of import controls with studied ambiguity. 2 As in his famous interview in Newsweek earlier this summer, he began by saying categorically that import controls were dangerous and damaging – 3 and then muddied the water by adding, "I do not rule out protective measures for particular industries suffering serious injury as a result of increased imports.
- 4 But if one thing is clear from the latest gloomy trade figures it is that import controls would not do much to stop the rot. 5 If one compares the past two three-month periods, imports appear to have risen quite sharply. 6 But the rise has not been in finished manufactures, the category where import controls would presumably be concentrated. 7 Even the volume of imported cars and other transport equipment fell between the two periods. 8 The increase has been concentrated in fuels, in chemicals, and in food, beverages and tobacco.
- 9 The fall in exports is more worrying than the rise in imports. 10 In the first half of this year, exports grew strongly. 11'Firms which at last found themselves with spare capacity on their hands managed to catch up on export backlogs, some of them stretching back shamefully far into 1973. 12 Now as recent surveys by the Confederation of British Industries have been revealing, new orders have petered out. 13 The problem is not basically one of price competitiveness-or so exporting firms report. 14 inflation has been largely offset by declining value of the pound abroad. 15 The trouble is quite simply the international slump. 16 In the past three months, only exports to the OPEC countries, where British firms have at last started lo do as well as their tougher competitors, have really increased strongly. 19 Exports to North America have fallen by 10 per cent.
- 18 To introduce "protective measures" now except in cases where dumping can be strictly proved under the rules of international trade, would be silly and self-defeating. 19 The governments in other industrial countries are under just the same sort of pressures as Mr. Wilson. 2o Last week, the US Treasury was told to start investigating the dumping of cars on the American markets and the protection lobby in the US is particularly worried about Volkswagen and British Leyland. 21 Curbs on Japanese car imports into Britain would make it harder for the US administration to resist curbs on British car imports into America. 22 Besides, Britain already has as tough a set of import controls as the Department of Trade would dare to dream up. 23 It is the £ 6-a-week pay policy. 24 With the fall in living standards that will take place this autumn, consumers are hardly going to have money to blow on imported luxuries.

Note

[1] I use "genre" in a broad sense that exceeds the idea of text and material and extends to practice as well. Some readers might prefer "discourse" or "discipline". "Discipline", however, invokes an institutional setting that "genre" does not. Though that setting is often present, the broader notion of genre often captures the stakes of my discussion better. Discourse, meanwhile, does not carry the emphasis on certain stylized elements of law that interest me in this essay. (Rosenberg 2014, p. 1057).

References

Ainsworth, Janet E. 2008. "You have the right to remain silent. . .But only if you ask for it just so": The role of linguistic ideology in American police interrogation law. *International Journal of Speech, Language and Law* 15: 1–21. [CrossRef]

Anesa, Patrizia, and Jan Engberg. 2023. *The Digital® Evolution of Legal Discourse. New Genres, Media and Linguistic Practices*. Berlin: De Gruyter Mouton.

Artemeva, Natasha, and Aviva Freedman. 2016. *Genre Studies Around the Globe*. Bloomington: Trafford Publications.

Bhatia, Vijay. 2023. Legal genres in interdiscursive contexts. In *Anne Wagner and Aleksandra Matulewska, Research Handbook on Jurilinguistics*. Northampton: Edward Elgar Publishing.

Biber, Douglas. 1995. *Dimensions of Register Variation: A Cross-Linguistic Comparison (CUP)*. Cambridge: Cambridge University Press.

Carston, Robyn, and Alison Hall. 2012. Implicature and explicature. In *Cognitive Pragmatics*. Edited by Hans-Jörg Schmi. Berlin: De Gruyter Mouton, pp. 47–84.

Chafe, Wallace. 1992. Immediacy and displacement in consciousness and language. In *Dieter Stein "Cooperating with Written Texts. The Pragmatics and Comprehension of Written Texts"*. Berlin: Mouton de Gruyter, pp. 231–56.

Dorgeloh, Heidrun, and Anja Wanner. 2022. *Discourse Syntax. English Grammar Beyond the Sentence*. Cambridge: Cambridge University Press.

Elder, Chi-Hé, and Kasia M. Jaszczolt. 2024. *Towards a Flexible Functional Proposition for Dynamic Discourse Meaning University of East Anglia, U.K.* Cambridge: University of Cambridge.

Engberg, Jan. 2013. Legal linguistics as a mutual arena for cooperation. Recent developments in the field of applied linguistics and law. *AILA Review* 26: 24–41. [CrossRef]

Giltrow, Janet. 2016. Form alone: The Supreme Court of Canada reading historical treaties. In *Genre Studies Around the Globe*. Edited by Natasha Artemeva and Aviva Freedman. Bloomington: Trafford Publications, pp. 207–24.

Giltrow, Janet, and Dieter Stein. 2009. Genres in the Internet: Innovation, evolution, and genre theory. In *Giltrow and Stein 2009*. Amsterdam: John Benjamins, pp. 1–26.

Greineder, Daniel, and Dieter Stein. 2023. The Internet as a game changer in legal communication: Arbitration on the move. *Anesa and Greineder* 2023: 85–96.

Guillén-Nieto, Victoria. 2024. *The Language of Harassment. Pragmatic Perspectives on Language as Evidence*. Marlborough: Lexington Books.

Guillén-Nieto, Victoria, Antonio Doval Pais, and Dieter Stein. 2023. *From Fear to Hate. Legal-Linguistic Perspectives on Migration*. Berlin: de Gruyter Mouton.

Jaszczolt, Kasia. 2011. Default meanings, salient meanings, and automatic processing. In *Salience and Defaults in Utterance Processing*. Edited by Kasia M. Jaszczolt and Keith Allan. Berlin: DeGruyter Mouton, pp. 11–34.

Jaszczolt, Kasia M. 2021. Functional proposition: A new concept for representing discourse meaning? *Journal of Pragmatics* 171: 200–14. [CrossRef]
Jaszczolt, Kasia M. 2023. *Semantics, Pragmatics, Philosophy. A Journey Through Meaning*. Cambridge: Cambridge University Press.
Kahneman, Daniel. 2012. *Thinking, Fast and Slow*. New York: Penguin.
Kuhlen, Anna Katharina, and Rasha Abdel Rahman. 2023. Beyond Speaking: Neurocognitive Perspectives on Language Production in Social Interaction. *Philosophical Transactions of the Royal Society B* 378: 20210483. [CrossRef]
Kurzon, Dennis. 1997. Legal Language: Varieties, Genres, Registers, Discourses. *International Journal of Applied Linguistics* 7: 119–39. [CrossRef]
Levinson, Stephen. 1992. Activity types and language. In *Talk at Work*. Edited by Paul Drew and John Heritage. Cambridge: Cambridge University Press, pp. 66–100.
Levinson, Stephen C. 2000. *Presumptive Meanings the Theory of Generalized Conversational Implicature*. Cambridge: MIT Press.
Levinson, Stephen C. 2024. *The Dark Matter of Pragmatics: Known Unknowns*. Cambridge: Cambridge University Press. [CrossRef]
Linderfalk, Ulf. 2007. *On the Interpretation of Treaties. The Modern International Law as Expressed in the 1969 Vienna Convention on the Law of Treaties*. Dordrecht: Springer.
Lyons, John. 1981. *Language, Meaning, and Context*. London: Fontana Paperbacks.
Makmillen, Shurli. 2007. Colonial texts in postcolonial contexts: A genre in the contact zone. *Linguistics and the Human Sciences* 3: 87–103. [CrossRef]
Nicklaus, Martina, and Dieter Stein. 2022. A lie or not a lie, that is the question. Trying to take arms against a sea of conceptual troubles: Methodological and theoretical issues in linguistic approaches to lie detection. In *Language as Evidence: Doing Forensic Linguistics*. Chp. 5. Cham: Palgrave Macmillan.
Palermo, Massimo, and Jacqueline Visconti. 2023. *Discourse Traditions, Text Linguistics and Historical Pragmatics*. Volume 10 of Manual of Discourse Traditions in Romance. Edited by Esme Winter-Froemel and Álvaro S. Octavio de Toledo y Huerta. Band 30 der Reihe Manuals of Romance Linguistics. Berlin: Walter de Gruyter, pp. 249–66. [CrossRef]
Puschmann, Cornelius. 2010. *The Corporate Blog as an Emerging Genre of Computer-Mediated Communication: Features, Constraints, Discourse Situation*. Göttingen: Universitätsverlag Göttingen.
Recanati, Francois. 2003. *Literal Meaning*. Cambridge: Cambridge University Press.
Richards, Ivor Armstrong. 1964. *The Philosophy of Rhetoric*. Oxford: Oxford University Press. First published 1937.
Rosenberg, Anat. 2014. The History of Genres: Reaching for Reality. *Law and Literature. Law & Social Inquiry* 39: 1057–79.
Rosenberg, Anat. 2022. *The Rise of Mass Advertising. Law, Enchantment, and the Cultural Boundaries of British Modernity*. Oxford: Oxford University Press.
Schaub, Thomas. 2010. Words are All we have: Text and Meaning in Pynchon's The crying of Lot 49. In *Text and Meaning. Literary Discourse and Beyond*. Edited by Richard Begam and Dieter Stein. Düsseldorf: Düsseldorf University Press, pp. 185–204.
Skoczen, Izabela. 2016. Minimal Semantics and Legal Interpretation. *Journal of the Semiotics of Law* 29: 615–33. [CrossRef]
Skoczen, Izabela. 2021. Implicatures within Legal Language—A Précis. Available online: https://researchgate.net/publication/355482995 (accessed on 30 May 2024).
Smolka, Jennifer, and Benedikt Pirker. 2016. International Law and Pragmatics—An account of Interpretation in International Law. *JLL* 5: 1–40.
Smolka, Jennifer, and Benedikt Pirker. 2018. International Law, Pragmatics and the Distinction Between Conceptual and Procedural Meaning. *JLL* 7: 117–41. [CrossRef]
Solan, Larry, and Peter Tiersma, eds. 2012. *The Oxford Handbook of Language and Law*. Oxford: Oxford University Press.
Stein, Dieter. 2022. Mobbing as a genre and cause for legal action? Linguistcprolegomena for a legal issue. *CORELA. Cognition, Représentation, Langage* H-36. [CrossRef]
Taboada, María Teresa. 2004. *Building Coherence and Cohesion. Task-Oriented Dialogue in English and Spanish*. Amsterdam: Benjamins.
Tiersma, Peter M. 2010. *Parchment, Paper, Pixels: Law and the Technologies of Communication*. Chicago: Chicago University Press.
Walton, Douglas, Fabrizio Macagno, and Giovanni Sartor. 2021. *Statutory Interpretation: Pragmatics and Argumentation*. Cambridge: Cambridge University Press.
Witczak-Plisiecka, Iwona. 2009. A note on the linguistic (in-)determinacy in the legal context. *Lodz Papers in Pragmatics* 5: 201–26. [CrossRef]

languages

MDPI

Article

The Language of Deception: Applying Findings on Opinion Spam to Legal and Forensic Discourses

Alibek Jakupov [1], Julien Longhi [2,*] and Besma Zeddini [1]

[1] SATIE Laboratory CNRS–UMR 8029, CY Tech, CY Cergy Paris University, 95000 Cergy, France; jakupovalibekdev@gmail.com (A.J.); besma.zeddini@cyu.fr (B.Z.)
[2] AGORA Laboratory EA 7392, CY Cergy Paris University, 95000 Cergy, France
* Correspondence: julien.longhi@cyu.fr

Abstract: Digital forensic investigations are becoming increasingly crucial in criminal investigations and civil litigations, especially in cases of corporate espionage and intellectual property theft as more communication occurs online via e-mail and social media. Deceptive opinion spam analysis is an emerging field of research that aims to detect and identify fraudulent reviews, comments, and other forms of deceptive online content. In this paper, we explore how the findings from this field may be relevant to forensic investigation, particularly the features that capture stylistic patterns and sentiments, which are psychologically relevant aspects of truthful and deceptive language. To assess these features' utility, we demonstrate the potential of our proposed approach using the real-world dataset from the Enron Email Corpus. Our findings suggest that deceptive opinion spam analysis may be a valuable tool for forensic investigators and legal professionals looking to identify and analyze deceptive behavior in online communication. By incorporating these techniques into their investigative and legal strategies, professionals can improve the accuracy and reliability of their findings, leading to more effective and just outcomes.

Keywords: digital investigation; NLP-based forensics; deceptive opinion spam; feature engineering; stylometry; sentiment analysis

Citation: Jakupov, Alibek, Julien Longhi and Besma Zeddini. 2024. The Language of Deception: Applying Findings on Opinion Spam to Legal and Forensic Discourses. *Languages* 9: 10. https://doi.org/10.3390/languages9010010

Academic Editor: Alan Garnham

Received: 16 October 2023
Revised: 10 December 2023
Accepted: 15 December 2023
Published: 22 December 2023

1. Introduction

Digital communication mediums like emails and social networks are crucial tools for sharing information and communication, but they can also be misused for criminal and political purposes. A notable instance of this misuse was the spread of false information during the U.S. election. Lazer et al. highlighted that "misinformation has become viral on social media" (Lazer et al. 2018). They underscored the importance for researchers and other relevant parties to encourage cross-disciplinary studies aimed at curbing the propagation of misinformation and addressing the root issues it exposes. Reports and worries have also arisen about terrorists and other criminal groups taking advantage of social media to promote their unlawful endeavors, such as setting up discrete communication pathways to share information (Goodman 2018). Therefore, it is not unexpected that government bodies are closely scrutinizing these platforms or communication paths. Most existing studies focus on creating a map of individual relationships within a communication network. The primary goal in these methods is to pinpoint the closest associates of a known target. These methods aim to enhance precision, recall, and/or the F1 score, often overlooking the significance of the content within conversations or messages. As a result, these methods can be highly specific (tailored for particular outcomes), may lack accuracy, and may not be ideal for digital investigations (Keatinge and Keen 2020). For example, in the tragic incident at the Gilroy Garlic Festival, the shooter had reportedly expressed his anger on his Facebook page before the incident. This post, however, did not attract the attention of pertinent parties until after the tragedy. This lack of attention is not surprising, given that

the shooter was not a recognized threat on the social network, and his post might not have been given high priority using traditional methods (Sun et al. 2021).

The example mentioned above demonstrates how written information can be employed to influence public opinion and impact the outcome of important events. There is a field within Natural Language Processing (NLP) that concentrates on scrutinizing on a similar phenomenon, called Deceptive Opinion Spam. Therefore, certain findings within this field could significantly enhance our comprehension of forensic linguistic analysis. Opinion Spam refers to reviews that are inappropriate or fraudulent, which can take on various forms such as self-promotion of an unrelated website or blog, or deliberate review fraud that could lead to monetary gain (Ott et al. 2011). Organizations have a strong incentive to detect and eliminate Opinion Spam via automation. This is because the primary concern with Opinion Spam is its influence on customer perception, particularly with regards to reviews that inaccurately praise substandard products or criticize superior ones (Vogler and Pearl 2020). Compared to other NLP tasks like sentiment analysis or intent detection, there has been relatively little research on using text classification approaches to detect Opinion Spam (Barsever et al. 2020). One can easily identify certain types of opinion spam, such as promotional content, inquiries, or other forms of non-opinionated text (Jindal and Liu 2008). The described situations can be classified as Disruptive Opinion Spam, characterized by irrelevant comments that are easily recognizable by the audience and pose a minimal threat, as individuals are empowered to disregard them if they so choose (Ott et al. 2011). When it comes to Deceptive Opinion Spam, which involves more nuanced forms of fake content, the task of identifying it is not as simple; the reason being that these statements are intentionally constructed to seem authentic and mislead the assessor (Ott et al. 2011). Deceptive Opinion Spam is a type of fraudulent behavior where a malicious user creates fictitious reviews, either positive or negative, with the intention of either boosting or damaging the reputation of a business or enterprise (Barsever et al. 2020). Thus, the deliberate intention to deceive readers in certain statements makes it challenging for human reviewers to accurately identify such deceptive texts, resulting in a success rate that is not significantly better than chance (Vogler and Pearl 2020). Consequently, discoveries in Deceptive Opinion Spam could prove valuable for designing digital investigation techniques for studying different communication channels, such as social networks. In contrast to traditional methods, the strategy that incorporates NLP techniques, particularly those used for Deceptive Opinion Spam analysis, places emphasis on both the interaction among individuals and the substance of the communication which may significantly improve the investigation process (Sun et al. 2021).

The problem is commonly addressed as a task of classifying text. Text classification systems typically consist of two key elements: a module for vectorization and a classifier. The vectorization module is tasked with creating features from a provided text sequence, while the classifier assigns category labels to the sequence using a set of matching features. These features are usually categorized into lexical and syntactic groups. Lexical features may include metrics such as total words or characters per word, as well as the frequency of long and unique words. On the other hand, syntactic features primarily consist of the frequency of function words or word groups, such as bag-of-words (BOW), n-grams, or Parts-Of-Speech (POS) tagging (Brown et al. 1992). In addition to vocabulary and sentence structure aspects, there are also methods known as lexicon containment techniques. These techniques symbolize the presence of a term from the lexicon in a text as a binary value, with positive indicating its existence and negative denoting its absence (Marin et al. 2014). The lexicons for such kind of features are constructed by a human expert (Pennebaker et al. 2001; Wilson et al. 2005) or generated automatically (Marin et al. 2010). Several approaches suggest integrating the text's morphological relationships and reliant linguistic components as input vectors for the classification algorithm (Brun and Hagege 2013). In addition to this, there are semantic vector space models which serve to characterize each word via a real-valued vector, determined using the distance or angle between pairs of word vectors (Sebastiani 2002). In the field of automatic fraudulent text

detection, various approaches have been applied, mostly relying on linguistic features, such as n-grams (Fornaciari and Poesio 2013; Mihalcea and Strapparava 2009; Ott et al. 2011), discourse structure (Rubin and Vashchilko 2012; Santos and Li 2009), semantically related keyword lists (Burgoon et al. 2003; Pérez-Rosas et al. 2015), measures of syntactic complexity (Pérez-Rosas et al. 2015), stylometric features (Burgoon et al. 2003), psychologically motivated keyword lists (Almela et al. 2015), and parts of speech (Fornaciari and Poesio 2014; Li et al. 2014).

These vectorization strategies are typically utilized to examine the significance of the features, which helps to highlight recurring patterns in the framework of fraudulent statements that are less prevalent in truthful texts. Although this technique shows some effectiveness, it has significant drawbacks due to the difficulty in controlling the quality of the training set. For example, while many of the classification algorithms, trained using this method, show acceptable performance within their specific fields, they struggle to generalize effectively across different domains, thereby lacking resilience in adapting to domain changes. (Krüger et al. 2017). As an illustration, a mere alteration in the polarity of fraudulent hotel evaluations (that is, training the model on positive reviews while testing it on negative ones) has the potential to significantly reduce the F score (Ott et al. 2013). This observation holds when the training and the testing dataset originate from different domains (Mihalcea and Strapparava 2009). Additionally, specific categorization models that rely on semantic vector space models could be significantly influenced by social or personal biases embedded in the training data. This can lead the algorithm to make incorrect deductions. (Papakyriakopoulos et al. 2020). Furthermore, certain studies suggest that deceptive statements differ from truthful ones more in terms of their sentiment then other linguistic features (Newman et al. 2003). According to certain cases, the deceivers display a more positive affect in order to mislead the audience (Zhou et al. 2004), whereas certain instances demonstrate that deception is characterized by more words reflecting negative emotion (Newman et al. 2003).

Based on the evidence mentioned above, it can be inferred that feature extraction methodologies utilized in classical NLP tasks exhibit limited reliability when applied to forensic investigations. This is primarily due to their strong association with particular lexical elements (like n-grams and specific keywords) or linguistically abstract components that may not be directly influenced by the style of verbal deception (such as specific parts of speech, stylometric features, and syntactic rules) (Vogler and Pearl 2020). From this point of view, it is more favorable to develop a novel set of features based on domain-independent approaches like sentiment analysis or stylometric features, as it offers superior generalization capabilities and independence from the training dataset domain.

2. Our Approach

Researchers in the forensic domain typically address investigative questions via linguistic analysis, such as identifying authors of illegal activities, understanding the content of documents, and extracting information about the timing, location, and intent of the text (Longhi 2021). Alternatively, studies into Deceptive Opinion Spam, which focus on fraudulent analysis, have proposed techniques for examining linguistic semantics by identifying patterns in the expression and content from a statistical standpoint. In fact, this method aligns with a forensic science approach, combining quantitative identification and qualitative analysis based on the analysis corpus consisting of different texts related to criminal acts, particularly involving terrorist groups, mostly in the same manner as scholars studying misleading discourse, but with the Ott Deceptive Opinion Spam corpus and the Multi-Domain Deceptive corpus instead (Jakupov et al. 2022). The goal is to assist investigators in finding stylistic similarities or exclusions between texts and potentially their authors.

In this paper, we explore the effectiveness of a novel linguistically defined implementation of stylometric and sentiment-based features for digital investigation. We begin by examining prior approaches to automatic fraudulent text detection, emphasizing tech-

niques that employ linguistic features such as n-grams, which provide the best performance within the domain. Following that, we outline the diverse corpora used to evaluate our approach and its cross-domain performance. Next, we explore the suggested sentiment-based features, confirming their possible significance in forensic examination within these collections. We also investigate the stylometric features and diagnostic potential of non-functional words, but without incorporating them into the classifier. Finally, we describe our classification scheme, which leverages these features.

2.1. Our Contributions

Our contributions can be summarized as follows.

- Novel approach to automatic digital forensic investigation that applies sentiment-based features;
- Comprehensive analysis of previous approaches to digital investigation, highlighting the strengths and weaknesses of different techniques and emphasizing the importance of linguistic features;
- Demonstration of the effectiveness of our approach using diverse corpora, showcasing its potential for forensic analysis;
- Investigation of the diagnostic potential of non-functional words as stylometric features

The significance of our contributions towards the advancement of automated digital forensic investigation lies in the incorporation of sentiment-based features, thereby transforming the paradigm of digital investigation methodologies. It particularly emphasizes the importance and diagnostic potential of non-functional words as stylometric features, which are typically overlooked by researchers.

Outline

The rest of the paper is organized as follows: in Section 3, we provide an overview of related work; in Section 4, we summarize our methodology for topic modeling and we present and discuss the experimental results as well as the datasets used to benchmark our approaches; finally, conclusions and discussions are provided in Sections 5 and 6.

3. Related Work

The idea of employing machine learning and deep learning methods to identify dubious activities in social networks has garnered general attention. For instance, Bindu et al. introduced an unsupervised learning method that can automatically spot unusual users in a static social network, albeit assuming that the network's structure does not change dynamically Bindu et al. (2017). Hassanpour et al. applied deep convolutional neural networks for images and long short-term memory (LSTM) to pull out predictive characteristics from Instagram's textual data, showing the capability to pinpoint potential substance use risk behaviors, aiding in risk evaluation and strategy formulation (Hassanpour et al. 2019). Tsikerdekis used machine learning to spot fraudulent accounts trying to enter an online sub-community for prevention purposes (Tsikerdekis 2016). Ruan et al. also used machine learning to detect hijacked accounts based on their online social behaviors (Ruan et al. 2015). Fazil and Abulaish suggested a mixed method to detect automated spammers on Twitter, using machine learning to examine related aspects like community-based features (e.g., metadata, content, and interaction-based features) (Fazil and Abulaish 2018). Cresci et al. employed machine learning to spot spammers using digital DNA technology, with the social fingerprinting technique designed to distinguish between spam bots and genuine accounts in both supervised and unsupervised manners (Cresci et al. 2017). Other applications focused on urban crime perception utilizing the convolutional neural network as their learning preference (Fu et al. 2018; Shams et al. 2018).

Certain studies showed the potential of focusing purely on textual data, especially in the context of social network analysis (Ala'M et al. 2017). One example of this application was in 2013, when Keretna et al. used a text mining tool, Stanford POS tagger, to pull out

features from Twitter posts that could indicate a user's specific writing style (Keretna et al. 2013). These features were then used in the creation of a learning module. Similarly, Lau et al. used both NLP and machine learning techniques to analyze Twitter data. They found that the Latent Dirichlet Allocation (LDA) and Support Vector Machine (SVM) methods yielded the best results in terms of the Area Under the ROC Curve (AUC) (Lau et al. 2014). In addition, Egele et al. developed a system to identify compromised social network accounts by analyzing message content and other associated features (Egele et al. 2015). Anwar and Abulaish introduced a unified social graph text mining framework for identifying digital evidence from chat logs based on user interaction and conversation data (Anwar and Abulaish 2014). Wang et al. treated each HTTP flow produced by mobile applications as text and used NLP to extract text-level features. These features were then used to create an effective malware detection model for Android viruses (Wang et al. 2017). Al-Zaidya et al. designed a method to efficiently find relevant information within large amounts of unstructured text data, visualizing criminal networks from documents found on a suspect's computer (Al-Zaidy et al. 2012). Lastly, Louis and Engelbrecht applied unsupervised information extraction techniques to analyze text data and uncover evidence, a method that could potentially find evidence overlooked by a simple keyword search (Louis and Engelbrecht 2011).

Li et al. applied their findings to detect fraudulent hotel reviews, using the Ott Deceptive Opinion spam corpus, and obtained a score of 81.8% by capturing the overall dissimilarities between truthful and deceptive texts (Li et al. 2014). The researchers expanded upon the Sparse Additive Generative Model (SAGE), which is a Bayesian generative model that combines both topic models and generalized additive models, and this resulted in the creation of multifaceted latent variable models via the summation of component vectors. Since most studies in this area focus on recognizing deceitful patterns instead of teaching a solitary dependable classifier, the primary difficulty of the research was to establish which characteristics have the most significant impact on each classification of a misleading review. Additionally, it was crucial to assess how these characteristics affect the ultimate judgment when they are paired with other attributes. SAGE is a suitable solution for meeting these requirements because it has an additive nature, which allows it to handle domain-specific attributes in cross-domain scenarios more effectively than other classifiers that may struggle with this task. The authors discovered that the BOW method was not as strong as LIWC and POS, which were modeled using SAGE. As a result, they formulated a general principle for identifying deceptive opinion spam using these domain-independent features. Moreover, unlike the creator of the corpus (Ott et al. 2011), they identified the lack of spatial information in hotel reviews as a potential indicator for identifying fraudulent patterns, of which the author's findings suggest that this methodology may not be universally appropriate since certain deceptive reviews could be authored by experts in the field. Although the research found that the domain-independent features were effective in identifying fake reviews with above-chance accuracy, it has also been shown that the sparsity of these features makes it difficult to utilize non-local discourse structures (Ren and Ji 2017); thus, the trained model may not be able to grasp the complete semantic meaning of a document. Furthermore, based on their findings, we can identify another significant indication of deceptive claims: the existence of sentiments. This is because reviewers often amplify their emotions by utilizing more vocabulary related to sentiments in their statements.

(Ren and Ji 2017) built upon earlier work by introducing a three-stage system. In the first stage, they utilized a convolutional neural network to generate sentence representations from word representations. This was performed by employing convolutional action, which is commonly used to synthesize lexical n-gram information. To accomplish this step, they employed three convolutional filters. These filters are effective at capturing the contextual meaning of n-grams, including unigrams, bigrams, and trigrams. This approach has previously proven successful for tasks such as sentiment classification. (Wilson et al. 2005). Subsequently, they created a model of the semantic and discourse relations of these sentence

vectors to build a document representation using a two-way gated recurrent neural network. These document vectors are ultimately utilized as characteristics to train a classification system. The authors achieved an 85.7% accuracy on the dataset created by Li et al. and showed that neural networks can be utilized to obtain ongoing document representations for the improved understanding of semantic features. The primary objective of this research was to practically show the superior efficacy of neural features compared to conventional discrete feature (like n-grams, POS, LIWC, etc.) due to their stronger generalization. Nevertheless, the authors' further tests showed that by combining discrete and neural characteristics, the total precision can be enhanced. Therefore, discrete features, such as the combination of sentiments or the use of non-functional words, continue to be a valuable reservoir of statistical and semantic data.

(Vogler and Pearl 2020) conducted a study investigating the use of particular details in identifying disinformation, both within a single area and across various areas. Their research focused on several linguistic aspects, including n-grams, POS, syntactic complexity metrics, syntactic configurations, lists of semantically connected keywords, stylometric properties, keyword lists inspired by psychology, discourse configurations, and named entities. However, they found these features to be insufficiently robust and adaptable, especially in cases where the area may substantially differ. This is mainly because most of these aspects heavily rely on specific lexical elements like n-grams or distinct keyword lists. Despite the presence of complex linguistic aspects such as stylometric features, POS, or syntactic rules, the researchers consider these to be of lesser importance because they do not stem from the psychological basis of verbal deceit. In their research, they saw deceit as a product of the imagination. Consequently, in addition to examining linguistic methods, they also explored approaches influenced by psychological elements, like information management theory (Burgoon et al. 1996), information manipulation theory (McCornack 1992), and reality monitoring and criteria-based statement analysis (Vogler and Pearl 2020). Since more abstract linguistic cues motivated by psychology may have wider applicability across various domains (Kleinberg et al. 2018), the authors find it beneficial to use these indicators grounded in psychological theories of human deception. They also lean on the research conducted by Krüger et al. which focuses on identifying subjectivity in news articles and proposes that linguistically abstract characteristics could potentially be more robust when used on texts from different fields (Krüger et al. 2017). For their experiment, Vogler and Pearl employed three different datasets for the purpose of training and evaluation, accommodating shifts in the domain, ranging from relatively subtle to considerably extensive: the Ott Deceptive Opinion Spam Corpus (Ott et al. 2011), essays on emotionally charged topics (Mihalcea and Strapparava 2009), and personal interview questions (Burgoon et al. 1996). The linguistically defined specific detail features the authors constructed for this research proved to be successful, particularly when there were notable differences in the domains used for training and testing. These elements were rooted in proper nouns, adjective phrases, modifiers in prepositional phrases, exact numeral terms, and noun modifiers appearing as successive sequences. The characteristics were derived from appropriate names, descriptive phrase clusters, prepositional phrase changes, precise numerical terms, and noun modifiers that showed up as successive sequences. Each attribute is depicted as the total normalized number and the average normalized weight. The highest F score they managed to obtain was 0.91 for instances where content remained consistent, and an F score of 0.64 for instances where there was a significant domain transition. This suggests that the linguistically determined specific detail attributes display a broader range of application. Even though the classifier trained with these features showed fewer false negatives, it struggled to accurately categorize truthful texts. The experimental results clearly indicate that a combination of n-gram and language-specific detail features tends to be more dependable only when a false positive carries a higher cost than a false negative. It is worth noting that features based on n-grams might have a superior ability for semantic expansion when they are built on distributed meaning representations like GloVe and ELMo. In their technique, however, n-gram

features rely only on single words without considering the semantic connection among them. This stands in stark contrast to our method, which revolves around analyzing the semantic essence of statements by evaluating the overall sentiment.

4. Materials and Methods

4.1. Model

Stylometry is a quantitative study of literary style that employs computational distant reading methods to analyze authorship. This approach is rooted in the fact that each writer possesses a distinctive, identifiable, and fairly stable writing style. This unique writing style is apparent in different writing components, including choice of words, sentence construction, punctuation, and the use of minor function words like conjunctions, prepositions, and articles. The fact that these function words are used unconsciously and independent of the topic makes them especially valuable for stylometric study.

In our research, we investigate the use of stylometric analysis in identifying misinformation, concentrating on the distinctive language patterns that can distinguish between honest and dishonest writings. Through the scrutiny of multiple stylometric aspects, our goal was to reveal the hidden features of dishonest language and establish a trustworthy approach for forensic investigation.

To obtain a better understanding of how lies are expressed in text, we utilized the Burrows' Delta method, a technique that gauges the "distance" between a text whose authorship is uncertain and another body of work. This approach is different from others like Kilgariff's chi-squared, as it is specifically structured to compare an unidentified text (or group of texts) with the signatures of numerous authors concurrently. More specifically, the Delta technique assesses how the unidentified text and groups of texts authored by an arbitrary number of known authors deviate from their collective average. Notably, the Delta method assigns equal importance to every characteristic it measures, thereby circumventing the issue of prevalent words dominating the outcomes, an issue often found in chi-squared tests. For these reasons, the Delta Method developed by John Burrows is typically a more efficient solution for authorship identification. We modified this method to discern the usage of non-functional words by deceivers and ordinary internet users. As this method extracts features that are not topic-dependent, we are able to establish a model that is resilient to changes in the domain.

Our adaptation of Burrows' original algorithm can be summarized as follows:

- Compile a comprehensive collection of written materials from a variable number of categories, which we will refer to as x (such as deceptive and truthful).
- Identify the top n words that appear most often in the dataset to utilize as attributes.
- For each of these n features, calculate the share of each of the x classes' subcorpora represented by this feature as a percentage of the total number of words. As an example, the word "the" may represent 4.72% of the words in the deceptive's subcorpus.
- Next, compute the average and standard deviation of these x values and adopt them as the definitive average and standard deviation for this characteristic across the entire body of work. Essentially, we will employ an average of the averages, rather than determining a sole value that symbolizes the proportion of the whole body of work represented by each term. We do this because we want to prevent a larger subsection of the body of work from disproportionately affecting the results and establish the standard for the body of work in a way that everything is presumed to resemble it.
- For each of the n features and x subcorpora, calculate a z score describing how far away from the corpus norm the usage of this particular feature in this particular subcorpus happens to be. To do this, subtract the "mean of means" for the feature from the feature's frequency in the subcorpus and divide the result by the feature's standard deviation. Below is the z-score equation for feature i, where $C(i)$ represents the observed frequency, the μ represents the mean of means, and the σ, the standard deviation.

$$Z_i = \frac{C_i - \mu_i}{\sigma_i} \quad (1)$$

- Next, calculate identical z scores for each characteristic in the text, where the authorship needs to be ascertained.
- Finally, compute a delta score to compare the unidentified text with each candidate's subset of text. This can be performed by calculating the mean of the absolute differences between the z scores for each characteristic in both the unidentified text and the candidate's text subset. This process ensures that equal weight is given to each feature, regardless of the frequency of words in the texts, preventing the top 3 or 4 features from overwhelming the others. The formula below presents the equation for Delta, where *Z(c,i)* represents the z score for feature *i* in candidate *c*, and *Z(t,i)* denotes the z score for feature *i* in the test case.

$$\Delta_c = \sum_i \frac{Z_c(i) - Z_t(i)}{n} \quad (2)$$

The class, or "winning" candidate, is most likely determined by finding the one with the least amount of difference in the score between their respective subcorpus and the test case. This indicates the least variation in writing style, which makes it the most probable class (either deceptive or truthful) for the text being examined.

In our methodology, we also incorporated a measure of exaggeration, consistently applied across various domains. The fundamental idea suggests that the intensity of the sentiment remains unchanged, irrespective of the text expressing a positive or negative sentiment (for instance, "I love the product" and "I detest the product" indicate the same level of sentiment, although in contrary directions). In order to examine false opinion spam, we made use of Azure Text Analytics API[1], which facilitates the analysis of the overall sentiment and the extraction of three aspects: positive, negative, and neutral. This was innately similar to the RGB color model, leading us to assign the values in the same way: Negative was paired with Red, Positive with Green, and Neutral with Blue. Following this, we displayed the pattern that began to form.

To illustrate the emotional trends in both honest and dishonest reviews, we initially utilized color-coding derived from sentiment analysis findings. To begin, we converted the sentiment ratings (positive, negative, and neutral) into a blue–green–red (BGR) format, which allowed us to represent each review as a pixel. Considering that Azure Text Analytics offers percentages for every sentiment component (e.g., 80% positive, 15% neutral, and 5% negative), we multiplied these values by 255 to facilitate visualization. Next, we devised auxiliary functions to convert sentiment scores into pixel format and generate an image utilizing the BGR values.

After recognizing visual patterns, we used these figures as attributes for our categorizer. To prevent the categorizer from making incorrect inferences by evaluating sentiments instead of hyperbole, we initially determined the total sentiment. If the sentiment was adverse, we exchanged the green and red channels, as hyperbole is steady for both negative and positive sentiments. We then standardized this set of attributes, as the percentage of neutral aspect is generally much higher than the other sentiments in most situations. Finally, we input these features into our classifier and examined the subsequent results as shown in Algorithm 1.

Algorithm 1 Extract Sentiment Features

$features \leftarrow []$)
for all $items \in Corpus$ **do**
  $sentiment \leftarrow mean(item.sentiments)$
  $aspect_{pos}, aspect_{neg}, aspect_{neut} \leftarrow item.sentiments$
  if $sentiment == Positive$ **then**
    $feature_r \leftarrow aspect_{neg} * 255$
    $feature_g \leftarrow aspect_{pos} * 255$
    $feature_b \leftarrow aspect_{neut} * 255$
  else
    $feature_r \leftarrow aspect_{pos} * 255$
    $feature_g \leftarrow aspect_{neg} * 255$
    $feature_b \leftarrow aspect_{neut} * 255$
  end if
  $feature \leftarrow (feature_r, feature_g, feature_b)$
  $feature \leftarrow normalize(feature)$
  $features \leftarrow feature$
end for

4.2. Data

Our initial approach involved examining labeled fraudulent reviews in order to train the model. One of the first large-scale, publicly available datasets for the research in this domain is Ott Deceptive Opinion Spam corpus (Ott et al. 2011), composed of 400 truthful and 400 gold-standard deceptive reviews. In order to obtain deceptive reviews of high quality via Amazon Mechanical Turk, a set of 400 Human-Intelligence Tasks (HITs) were created and distributed among 20 selected hotels. To ensure uniqueness, only one submission per Turker was allowed. To obtain truthful reviews, the authors gathered 6977 reviews from the 20 most popular Chicago hotels on Trip Advisor. Despite the dataset, the authors have discovered that detecting deception is a challenge for human judges, as most of them performed poorly.

To prevent our model from identifying inaccurate features that are related to the domain rather than deceptive cues, we augmented our training dataset with cross-domain data. For cross-domain investigation, we applied a dataset consisting of hotel, restaurant, and doctor reviews (Li et al. 2014) obtained from various sources, including TripAdvisor and Amazon. The deceptive reviews were primarily procured from two sources: professional content writers and participants from Amazon Mechanical Turk. This approach allowed the researchers to capture the nuances of deceptive opinions generated by both skilled and amateur writers. To ensure the quality and authenticity of truthful reviews, the authors relied on reviews with a high number of helpful votes from other users. This criterion established a baseline of credibility for the truthful reviews in the dataset. Furthermore, the dataset included reviews with varying sentiment polarities (positive and negative) to account for the sentiment intensity and exaggeration aspects in deceptive opinion spam.

Following the model's training, we opted to assess its usefulness in forensic investigations by evaluating it on real-world email data. Email serves as a crucial means of communication within most businesses, facilitating internal dialogue between staff members and external communication with the broader world. Consequently, it offers a wealth of data that could potentially highlight issues. However, this brings up the issue of privacy, as the majority of employees would not be comfortable knowing their employer has access to their emails. Therefore, it is critical to adopt methods to manage this issue that are as non-invasive as possible. This is also beneficial to the organization, as implementing a system that literally "reads" employees' emails could prove to be excessively costly.

Theories of deceptive behavior, fraud, or conspiracy suggest that changes in language use can signal elements such as feelings of guilt or self-awareness regarding the deceit, as well as a reduction in complexity to ease the consistency of repetition and lessen the mental load of fabricating a false narrative (Keila and Skillicorn 2005). The potential presence

of some form of monitoring may also lead to an excessive simplicity in messages, as the senders strive to avoid detection. This simplicity could, in itself, become a telltale sign. It is also probable that messages exchanged between collaborators will contain abnormal content, given that they are discussing actions that are unusual within their context.

The Enron email dataset was made publicly available in 2002 by the Federal Energy Regulatory Commission (FERC). This dataset consists of real-world emails that were sent and received by ex-Enron employees. The dataset contains 517,431 emails from the mail folders of 150 ex-Enron employees, including top executives such as Kenneth Lay and Jeffrey Skilling. While most of the communication in the dataset is mundane, some emails from executives who are currently being prosecuted suggest the presence of deceptive practices. The emails contain information such as sender and receiver email addresses, date, time, subject, body, and text, but do not include attachments. This dataset is widely used for research purposes and was compiled by Cohen at Carnegie Mellon University. We initiated a preprocessing phase to polish the dataset, which involved eliminating redundant entries, junk emails, unsuccessful and blank emails, along with punctuation symbols (essential for applying sentiment analysis). This purification process resulted in a remaining total of 47,468 emails, all of which were either dispatched or obtained by 166 previous Enron employees. Among these employees, 25 were marked as "criminals", a term denoting those who were supposedly involved in fraudulent acts.

5. Results

At first, we analyzed a group of deceptive reviews which consisted of the Ott Deceptive Opinion Spam Corpus and the cross-domain corpus of reviews for hotels, restaurants, and doctors curated by Li et al. Our aim was to confirm that the use of non-essential words remained consistent across various domains. The combined dataset was divided into a 25% test set and a 75% training set, and the training set was used to evaluate the accuracy of correct identification. The results of the negative deceptive test indicated a delta score of 1.3815 for deceptive and 1.8281 for truthful, while the negative truthful test had a delta score of 1.4276 for deceptive and 1.0704 for truthful. As for the positive tests, the deceptive test had a delta score of 1.4003 for deceptive and 1.8459 for truthful, whereas the truthful test had a delta score of 2.9074 for deceptive and 2.2098 for truthful. Overall, the model accurately detected 65% of deceptive texts and 68% of truthful texts, taking into account both positive and negative cases.

The study primarily investigated the stylometric characteristics and potential usefulness of non-functional words, but decided not to include them in the classifier due to the inherent methodological limitation that necessitates analyzing the entire corpus for vectorizing individual statements. However, the results uncovered interesting patterns that require further exploration and may be potentially applied to forensic investigation.

After exploring the fraudulent reviews, we focused on extracting sentiment-based features. To observe emotional trends in truthful and deceptive reviews, we colored the reviews using a blue–green–red (BGR) format based on their sentiment scores (positive, negative, and neutral). This allowed us to depict each review as a pixel, with blue indicating neutral sentiment, green representing positive sentiment, and red signifying negative sentiment. To convert the sentiment scores into pixel format and create an image from the BGR values, we developed support functions. Each image showcased 400 pixels (20×20), symbolizing 400 reviews.

We created images for different categories of reviews, such as deceptive positive, deceptive negative, truthful positive, and truthful negative, and compared their visual patterns. The analysis showed that fake negative reviews had a brighter appearance with less green spots, whereas fake positive reviews had more vibrant colors with fewer red spots. This suggests that there is an element of exaggeration and insincere praise in deceitful reviews. Conversely, truthful reviews appeared to be more authentic and impartial in their emotional tone.

In order to achieve a consistent color that conveys deception, we took all the pixels in the images and computed their average values across three color channels: blue, green, and red. Afterward, we combined the channels to create a single color that symbolizes the mean sentiment of the dishonest reviews, as shown in Figure 1.

According to the study, negative reviews that were truthful appeared to be less red in color than negative reviews that were deceptive. On the other hand, positive reviews that were fake appeared to be greener than positive reviews that were truthful. This indicates that deceptive reviews tend to contain more exaggerated expressions of sentiment, which can be represented through the use of color.

Figure 1. Deceptive datasets: colorized sentiments.

With this in mind, we trained multiple classifiers with features extracted using Algorithm 1. The training was conducted with the Ott Deceptive Opinion Spam dataset, while the Li et al. cross-domain dataset was used for testing. Once we identified the optimal model, we applied it to the Enron email corpus.

In order to ensure that the input features used in a machine learning model have a consistent scale or distribution, we applied different normalization techniques such as MaxAbsScaler, StandardScaler Wrapper, and Sparse Normalizer in our experiment. We chose AUC Weighted as the primary metric to assess the performance of our models. AUC Weighted was selected because it is capable of measuring the classifier's performance across varying thresholds, while also considering the potential class imbalance present in the cross-domain dataset. This guarantees a more reliable and strong evaluation of the model's ability to differentiate truthful and deceptive opinions.

Table 1 clearly indicates that the classifier's performance is consistent, signifying that the features are robust even in cross-domain situations. It should be emphasized that the merged dataset encompasses various fields and includes both favorable and unfavorable evaluations. This implies that the suggested characteristics can proficiently endure changes in the sentiment as well.

Table 1. Classifiers utilizing sentiment-based features

Algorithm	Normalizer	AUC Weighted
Light GBM	Sparse Normalizer	0.67
Random Forest	Sparse Normalizer	0.68
Light GBM	Standard Scaler Wrapper	0.68
Light GBM	Max Abs Scaler	0.69
Random Forest	Max Abs Scaler	0.69
Random Forest	Standard Scaler Wrapper	0.70
Logistic Regression	Standard Scaler Wrapper	0.71
Extreme RandomTrees	Max Abs Scaler	0.73
Light GBM	Standard Scaler Wrapper	0.74
Extreme Random Trees	Max AbsScaler	0.74

While there is a reduction in accuracy compared to related work, we can still achieve relatively high and stable results, which is more important since it reduces the risk of

overfitting. Our progress in this area is leading us towards developing a universal method for detecting deception, rather than creating a classifier that is only suitable for a particular dataset. This approach proves to be more effective in identifying instances of deception on the internet.

The model trained on the deceptive training set was finally applied to the Enron email dataset, including mails from high-ranking executives like Kenneth Lay (ex-Chairman and CEO) and Jeffrey Skilling (ex-CEO). Although the majority of the communication is innocuous and uneventful, the emails of several executives who are currently facing prosecution are included in the dataset, suggesting that evidence of deception could potentially be found within the data. We cross-referenced the name list on the website to confirm the authenticity of the email and determine whether it is misleading. Our model was able to obtain the F1 score of 0.43, but due to the dataset being imbalanced, with only 25 out of 166 employees being identified as criminals, our evaluation of the model takes into account some level of uncertainty.

In order to comprehend how our model can be applied in practical scenarios, we assessed its performance against other top-performing models such as SIIMCO (Taha and Yoo 2016) and LogAnalysis (Ferrara et al. 2014), despite them not being rooted in NLP. These methods were devised by building an extensive graph detailing the suspected individuals' connections, with those particularly active in the communication network frequently being strongly implicated as criminals. For example, "employee 57", who exchanged 3247 and 847 emails, respectively, was identified as a criminal as per both existing techniques, or in other words, a true negative.

Upon examining Table 2, it is clear that our approach yields a lower F1 score and precision rate. This disparity can be attributed to several factors.

Firstly, our classifier was trained exclusively on online reviews, excluding emails or any other communication types involving two or more parties. This specificity could affect the textual patterns we can detect. As a result, it would be beneficial to enrich our training set with anonymized conversation data.

Secondly, our preprocessing stage overlooked the removal of email signatures and conversation history. This oversight could distort the analysis results, as the response may not be deceptive itself, but it could contain traces of a previous deceptive email. Consequently, we must refine our text preprocessing pipeline and integrate a layout analysis to distinguish the message body from the metadata, such as signatures or conversation history.

Lastly, the level of exaggeration, which is commonplace in online reviews, may not translate accurately to the corporate communication realm. Therefore, we should consider introducing a variable exaggeration level that adapts to the specific domain.

Table 2. Performance of SIIMCO and LogAnalysis: A comparative summary.

Approach	F1 Score	Precision	Recall
LogAnalysis	0.51	0.49	0.53
SIIMCO	0.59	0.58	0.60
Our proposed approach	0.43	0.26	1

6. Discussion

Current state-of-the-art models, based on common features like n-grams or embeddings, have demonstrated their effectiveness within specific domains, with improvements achieved when combined with other features. However, cross-domain performance tends to decrease as content differences between training and testing datasets increase. The utilization of more abstract linguistic features, such as syntax-based features and psychologically motivated categories, has shown to enhance cross-domain deception detection performance.

Our method has been shown to be effective in detecting deception in various deceptive reviews. Stylometric analysis, which focuses on unique linguistic patterns in writing, has

demonstrated promise in uncovering the underlying characteristics of deceptive language. Sentiment analysis and visualization techniques have also been explored to identify patterns in deceptive and truthful reviews. Converting sentiment scores into color formats and generating images to represent reviews allows for visual comparison and insights into exaggeration levels present in online communication.

However, for better performance on email data, like the Enron dataset, one alternative approach we could have used is a transductive method, specifically by employing topic modeling, such as the LDA model, on the entire dataset. Moreover, we would recommend evaluating the model using a 5×2 Nested Cross Validation method. This involves splitting the preprocessed dataset into five folds, with each fold potentially being chosen as the test set, while the remaining four are used for a 2-fold validation. The training set should then be used to train the classifier, with each generator building a group of classifiers for each possible number of topics from zero up to the number given by the LDA, with the smallest perplexity. The validation set should be used to test these classifiers in terms of precision, recall, and F1 score. Only the best classifiers for each metric should be recommended to the investigator and evaluated in the test set.

To sum up, the insights gained from studying the linguistic and psychological aspects of deception can be leveraged to improve existing tools used by investigators and legal professionals tasked with identifying deceptive behavior in online communication. By providing these individuals with a deeper understanding of the subtle markers that indicate deception, they may be better equipped to assess the credibility of information and make informed decisions in high-stakes situations.

7. Conclusions

The results of our study have significant implications for cross-domain approaches in the future and we have specific suggestions. Firstly, it should be expected that there will be a decline in classification performance when transitioning from within-domain to cross-domain detection, regardless of the approach used. Our study has investigated specific details in this regard, but they are unable to completely negate this drop in performance. Therefore, if possible, it is recommended to use training data that is closely related to the testing data in terms of domain, with a closer match being preferable.

However, when this is not feasible, and the training content differs significantly from the test content, it is important to weigh the tradeoff between false negatives and false positives. If false negatives are a greater concern, relying solely on linguistically defined specific details can be advantageous. On the other hand, if false positives are the greater concern, it is preferable to use a combination of n-gram and linguistically defined specific detail features.

Our study draws on insights from prior deception detection methods, including both within-domain and cross-domain approaches, to identify linguistically defined sentiment and stylometric features that can effectively be applied for forensic investigation across domains under specific circumstances. These features are particularly useful when there are significant content differences between training and test sets, as well as when the cost of false negatives is greater than that of false positives. We anticipate that future research will use these findings to improve general-purpose forensic investigation strategies.

In essence, the advancements made in the field of Deceptive Opinion Spam detection not only hold the potential to improve trust and transparency in online communications, but also contribute to the broader domains of online threat investigation. As research in this area continues to evolve, it is crucial that the knowledge and methodologies developed are shared and adapted across disciplines, thereby maximizing their impact and benefit to society as a whole.

Author Contributions: Conceptualization, A.J. and J.L.; methodology, A.J. and J.L.; software, A.J.; validation, J.L. and B.Z.; formal analysis, B.Z.; investigation, A.J.; resources, J.L.; data curation, A.J.; writing—original draft preparation, A.J.; writing—review and editing, J.L.; visualization, A.J.;

supervision, J.L.; project administration, B.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data supporting the findings of this study are available in the author's GitHub repository: https://github.com/ajakupov/ColorizeComments (accessed on 7 July 2023), as well as on the following websites: https://myleott.com/op-spam.html, https://nlp.stanford.edu/~bdlijiwei/Code.html (accessed on 1 June 2023) and https://www.cs.cmu.edu/~enron/ (accessed on 20 August 2023).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

NLP	Natural Language Processing
BOW	Bag of Words
POS	Part of Speech
LSTM	Long Short-Term Memory Networks
DNA	Deoxyribonucleic Acid
LDA	Latent Dirichlet Allocation
SVM	Support Vector Machine
FERC	Federal Energy Regulatory Commission
ROC	Rate of Change
AUC	Area under the ROC Curve
HTTP	Hypertext Transfer Protocol
SAGE	Sparse Additive Generative Model
LIWC	Linguistic Inquiry and Word Count
GloVe	Global Vectors for Word Representation
ELMo	Embeddings from Language Model

Note

[1] https://learn.microsoft.com/en-us/azure/cognitive-services/language-service/sentiment-opinion-mining/overview, accessed on 21 September 2023.

References

Al-Zaidy, Rabeah, Benjamin C. M. Fung, Amr M. Youssef, and Francis Fortin. 2012. Mining criminal networks from unstructured text documents. *Digital Investigation* 8: 147–60. [CrossRef]

Ala'M, Al-Zoubi, Ja'far Alqatawna, and Hossam Paris. 2017. Spam profile detection in social networks based on public features. Paper presented at 2017 8th International Conference on information and Communication Systems (ICICS), Irbid, Jordan, April 4–6. pp. 130–35.

Almela, Ángela, Gema Alcaraz-Mármol, and Pascual Cantos. 2015. Analysing deception in a psychopath's speech: A quantitative approach. *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada* 31: 559–72. [CrossRef]

Anwar, Tarique, and Muhammad Abulaish. 2014. A social graph based text mining framework for chat log investigation. *Digital Investigation* 11: 349–62. [CrossRef]

Barsever, Dan, Sameer Singh, and Emre Neftci. 2020. Building a better lie detector with bert: The difference between truth and lies. Paper presented at 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, July 19–24. pp. 1–7.

Bindu, P. V., P. Santhi Thilagam, and Deepesh Ahuja. 2017. Discovering suspicious behavior in multilayer social networks. *Computers in Human Behavior* 73: 568–82. [CrossRef]

Brown, Peter F., Vincent J. Della Pietra, Peter V. Desouza, Jennifer C. Lai, and Robert L. Mercer. 1992. Class-based n-gram models of natural language. *Computational Linguistics* 18: 467–80.

Brun, Caroline, and Caroline Hagege. 2013. Suggestion mining: Detecting suggestions for improvement in users' comments. *Research in Computing Science* 70: 5379–62. [CrossRef]

Burgoon, Judee K., David B. Buller, Laura K. Guerrero, Walid A. Afifi, and Clyde M. Feldman. 1996. Interpersonal deception: Xii. information management dimensions underlying deceptive and truthful messages. *Communications Monographs* 63: 50–69. [CrossRef]

Burgoon, Judee K., J. Pete Blair, Tiantian Qin, and Jay F. Nunamaker. 2003. Detecting deception through linguistic analysis. Paper presented at Intelligence and Security Informatics: First NSF/NIJ Symposium, ISI 2003, Tucson, AZ, USA, June 2–3; Proceedings 1, Berlin/Heidelberg: Springer, pp. 91–101.

Cresci, Stefano, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. 2017. Social fingerprinting: Detection of spambot groups through dna-inspired behavioral modeling. *IEEE Transactions on Dependable and Secure Computing* 15: 561–76. [CrossRef]

Egele, Manuel, Gianluca Stringhini, Christopher Kruegel, and Giovanni Vigna. 2015. Towards detecting compromised accounts on social networks. *IEEE Transactions on Dependable and Secure Computing* 14: 447–60. [CrossRef]

Fazil, Mohd, and Muhammad Abulaish. 2018. A hybrid approach for detecting automated spammers in twitter. *IEEE Transactions on Information Forensics and Security* 13: 2707–19. [CrossRef]

Ferrara, Emilio, Pasquale De Meo, Salvatore Catanese, and Giacomo Fiumara. 2014. Detecting criminal organizations in mobile phone networks. *Expert Systems with Applications* 41: 5733–50. [CrossRef]

Fornaciari, Tommaso, and Massimo Poesio. 2013. Automatic deception detection in italian court cases. *Artificial Intelligence and Law* 21: 303–40. [CrossRef]

Fornaciari, Tommaso, and Massimo Poesio. 2014. Identifying fake amazon reviews as learning from crowds. Paper presented at 14th Conference of the European Chapter of the Association for Computational Linguistics, Gothenburg, Sweden, April 26–30. Toronto: Association for Computational Linguistics: pp. 279–87.

Fu, Kaiqun, Zhiqian Chen, and Chang-Tien Lu. 2018. Streetnet: Preference learning with convolutional neural network on urban crime perception. Paper presented at 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Seattle, WA, USA, November 6–9. pp. 269–78.

Goodman, Anka Elisabeth Jayne. 2018. When you give a terrorist a twitter: Holding social media companies liable for their support of terrorism. *Pepperdine Law Review* 46: 147.

Hassanpour, Saeed, Naofumi Tomita, Timothy DeLise, Benjamin Crosier, and Lisa A. Marsch. 2019. Identifying substance use risk based on deep neural networks and instagram social media data. *Neuropsychopharmacology* 44: 487–94. [CrossRef]

Jakupov, Alibek, Julien Mercadal, Besma Zeddini, and Julien Longhi. 2022. Analyzing deceptive opinion spam patterns: The topic modeling approach. Paper presented at 2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI), Macao, China, October 31–November 2, pp. 1251–61.

Jindal, Nitin, and Bing Liu. 2008. Opinion spam and analysis. Paper presented at 2008 International Conference on Web Search and Data Mining, Palo Alto, CA, USA, February 11–12. pp. 219–30.

Keatinge, Tom, and Florence Keen. 2020. Social media and (counter) terrorist finance: A fund-raising and disruption tool. In *Islamic State's Online Activity and Responses*. London: Routledge, pp. 178–205.

Keila, Parambir S., and David B. Skillicorn. 2005. Detecting unusual and deceptive communication in email. Paper presented at Centers for Advanced Studies Conference, Toronto, ON, Canada, October 17–20. Pittsburgh: Citeseer, pp. 17–20.

Keretna, Sara, Ahmad Hossny, and Doug Creighton. 2013. Recognising user identity in twitter social networks via text mining. Paper presented at 2013 IEEE International Conference on Systems, Man, and Cybernetics, Manchester, UK, October 13–16. pp. 3079–82.

Kleinberg, Bennett, Maximilian Mozes, Arnoud Arntz, and Bruno Verschuere. 2018. Using named entities for computer-automated verbal deception detection. *Journal of Forensic Sciences* 63: 714–23. [CrossRef]

Krüger, Katarina R., Anna Lukowiak, Jonathan Sonntag, Saskia Warzecha, and Manfred Stede. 2017. Classifying news versus opinions in newspapers: Linguistic features for domain independence. *Natural Language Engineering* 23: 687–707. [CrossRef]

Lau, Raymond Y. K., Yunqing Xia, and Yunming Ye. 2014. A probabilistic generative model for mining cybercriminal networks from online social media. *IEEE Computational Intelligence Magazine* 9: 31–43. [CrossRef]

Lazer, David M. J., Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, and et al. 2018. The science of fake news. *Science* 359: 1094–96. [CrossRef]

Li, Jiwei, Myle Ott, Claire Cardie, and Eduard Hovy. 2014. Towards a general rule for identifying deceptive opinion spam. Paper presented at 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Baltimore, MD, USA, June 22–27. pp. 1566–76.

Longhi, Julien. 2021. Using digital humanities and linguistics to help with terrorism investigations. *Forensic Science International* 318: 110564. [CrossRef]

Louis, A. L., and Andries P. Engelbrecht. 2011. Unsupervised discovery of relations for analysis of textual data. *Digital Investigation* 7: 154–71. [CrossRef]

Marin, Alex, Mari Ostendorf, Bin Zhang, Jonathan T. Morgan, Meghan Oxley, Mark Zachry, and Emily M. Bender. 2010. Detecting authority bids in online discussions. Paper presented at 2010 IEEE Spoken Language Technology Workshop, Berkeley, CA, USA, December 12–15. pp. 49–54.

Marin, Alex, Roman Holenstein, Ruhi Sarikaya, and Mari Ostendorf. 2014. Learning phrase patterns for text classification using a knowledge graph and unlabeled data. Paper presented at Fifteenth Annual Conference of the International Speech Communication Association, Singapore, September 14–18.

McCornack, Steven A. 1992. Information manipulation theory. *Communications Monographs* 59: 1–16. [CrossRef]
Mihalcea, Rada, and Carlo Strapparava. 2009. The lie detector: Explorations in the automatic recognition of deceptive language. Paper presented at ACL-IJCNLP 2009 Conference Short Papers, Singapore, August 4, pp. 309–12.
Newman, Matthew L., James W. Pennebaker, Diane S. Berry, and Jane M. Richards. 2003. Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin* 29: 665–75. [CrossRef]
Ott, Myle, Claire Cardie, and Jeffrey T. Hancock. 2013. Negative deceptive opinion spam. Paper presented at 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Atlanta, Georgia, June 9–14. pp. 497–501.
Ott, Myle, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. *arXiv* arXiv:1107.4557.
Papakyriakopoulos, Orestis, Simon Hegelich, Juan Carlos Medina Serrano, and Fabienne Marco. 2020. Bias in word embeddings. Paper presented at 2020 Conference on Fairness, Accountability, and Transparency,Barcelona, Spain, January 27–30. pp. 446–57.
Pennebaker, James W., Martha E. Francis, and Roger J. Booth. 2001. *Linguistic Inquiry and Word Count: Liwc 2001*. Mahway: Lawrence Erlbaum Associates, vol. 71.
Pérez-Rosas, Verónica, Mohamed Abouelenien, Rada Mihalcea, and Mihai Burzo. 2015. Deception detection using real-life trial data. Paper presented at 2015 ACM on International Conference on Multimodal Interaction, Seattle, WA, USA, November 9–13. pp. 59–66.
Ren, Yafeng, and Donghong Ji. 2017. Neural networks for deceptive opinion spam detection: An empirical study. *Information Sciences* 385: 213–24. [CrossRef]
Ruan, Xin, Zhenyu Wu, Haining Wang, and Sushil Jajodia. 2015. Profiling online social behaviors for compromised account detection. *IEEE Transactions on Information Forensics and Security* 11: 176–87. [CrossRef]
Rubin, Victoria L., and Tatiana Vashchilko. 2012. Identification of truth and deception in text: Application of vector space model to rhetorical structure theory. Paper presented at Workshop on Computational Approaches to Deception Detection, Avignon, France, April 23; pp. 97–106.
Santos, Eugene, and Deqing Li. 2009. On deception detection in multiagent systems. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 40: 224–35. [CrossRef]
Sebastiani, Fabrizio. 2002. Machine learning in automated text categorization. *ACM Computing Surveys (CSUR)* 34: 1–47. [CrossRef]
Shams, Shayan, Sayan Goswami, Kisung Lee, Seungwon Yang, and Seung-Jong Park. 2018. Towards distributed cyberinfrastructure for smart cities using big data and deep learning technologies. Paper presented at 2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS), Vienna, Austria, July 2–6. pp. 1276–83.
Sun, Dongming, Xiaolu Zhang, Kim-Kwang Raymond Choo, Liang Hu, and Feng Wang. 2021. Nlp-based digital forensic investigation platform for online communications. *Computers & Security* 104: 102210.
Taha, Kamal, and Paul D. Yoo. 2016. Using the spanning tree of a criminal network for identifying its leaders. *IEEE Transactions on Information Forensics and Security* 12: 445–53. [CrossRef]
Tsikerdekis, Michail. 2016. Identity deception prevention using common contribution network data. *IEEE Transactions on Information Forensics and Security* 12: 188–99. [CrossRef]
Vogler, Nikolai, and Lisa Pearl. 2020. Using linguistically defined specific details to detect deception across domains. *Natural Language Engineering* 26: 349–73. [CrossRef]
Wang, Shanshan, Qiben Yan, Zhenxiang Chen, Bo Yang, Chuan Zhao, and Mauro Conti. 2017. Detecting android malware leveraging text semantics of network flows. *IEEE Transactions on Information Forensics and Security* 13: 1096–1109. [CrossRef]
Wilson, Theresa, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. Paper presented at Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Vancouver, BC, Canada, October 6–8. pp. 347–54.
Zhou, Lina, Judee K. Burgoon, Douglas P. Twitchell, Tiantian Qin, and Jay F. Nunamaker, Jr. 2004. A comparison of classification methods for predicting deception in computer-mediated communication. *Journal of Management Information Systems* 20: 139–66. [CrossRef]

languages

MDPI

Article

Beyond "I Didn't Do It": A Linguistic Analysis of Denial in US Legal Settings

Elena Didoni [1] and Claudia Roberta Combei [2,*]

[1] Laboratorio di Neurolinguistica e Pragmatica Sperimentale (NEP), Dipartimento di Scienze Umane e della Vita, Istituto Universitario di Studi Superiori (IUSS), 27100 Pavia, Italy; elena.didoni@iusspavia.it
[2] Dipartimento di Studi Umanistici, Università di Pavia, 27100 Pavia, Italy
* Correspondence: claudiaroberta.combei@unipv.it

Abstract: This paper investigates the multimodal manifestations of denial in US legal contexts, in the English language, by analyzing police interviews and cross-examinations. The research uses a 13-h corpus of video recordings portraying five male suspects, aged 20–44, eventually charged and convicted of femicide. We deploy techniques from conversation analysis, multimodal analysis, and speech processing, using tools like ELAN, Praat, WebMAUS, and Python libraries to transcribe, annotate, and analyze audio–video data. This exploratory study identifies several recurring patterns in prosodic and gestural cues associated with denial. In particular, our results indicate a prototypical multimodal denial characterized by a predominant gestural component: head positioning (neutral or lowered) and head shaking. This gestural expression is frequently repeated and can also function independently as a nonverbal marker of denial. Denial is also often accompanied by open-hand gestures, sitting upright posture, and a certain degree of vagueness in speech. Furthermore, our findings suggest that the expression of denial often involves a reduction in pitch and intensity following the confession or indictment. The analysis of pauses before denial instances reveals that a greater number of pauses typically occurs after incrimination. Overall, this study shows that there is an interesting interplay between verbal and nonverbal features of denial in legal interactions, underscoring the need for further analysis.

Keywords: denial; multimodality; forensic linguistics

Citation: Didoni, Elena, and Claudia Roberta Combei. 2024. Beyond "I Didn't Do It": A Linguistic Analysis of Denial in US Legal Settings. *Languages* 9: 351. https://doi.org/10.3390/languages9110351

Academic Editors: Julien Longhi and Nadia Makouar

Received: 31 July 2024
Revised: 1 October 2024
Accepted: 11 November 2024
Published: 19 November 2024

Correction Statement: This article has been republished with a minor change. The change does not affect the scientific content of the article and further details are available within the backmatter of the website version of this article.

1. Introduction

Negation represents a fundamental aspect of human language that is rooted in cognitive processes. The expression of negation begins in early childhood (Morris 2003) and, as Horn (2010) claims, it is an essential device of the communicative system, since it furnishes speakers with the tools for denial, contradiction, misrepresentation, deception, and irony. Abandoning the simplistic view of negation as a mere binary operator that assigns truth values, recent research has described negation as a complex cognitive, linguistic, and logical device displaying complex syntactic, semantic, and pragmatic properties and functions (Prieto and Espinal 2020).

Indeed, languages possess morphological, syntactic, and semantic mechanisms that allow speakers to express negation verbally. In face-to-face communication, verbal expressions of negation are frequently supplemented by nonverbal cues, such as prosody (e.g., intensity, pitch, etc.) and gestural behavior (e.g., hand gestures, shoulder shrugs, etc.). These nonverbal devices can also operate independently as, for instance, head shaking is associated with negation in certain cultural and linguistic contexts.

The evolution of negation spans from the basic act of refusal, a communicative behavior that is already present in early stages of language development and is shared with animals, to a sophisticated range of conceptually grounded uses exclusive to human beings. Actually, negation serves a variety of communicative purposes, including the expression

of falsity, absence, non-existence, denial, rejection, and correction (Vandamme 1972). In this respect, Roitman (2017) refines the perspective by asserting that linguistic negation encompasses three primary meanings: non-existence, rejection, and denial.

The scope of this paper is to examine one of these three specific functions, namely 'denial'. We will focus on the distinctive realizations of denial analyzed from a multimodal perspective (i.e., gestures and prosody). Pragmatic aspects, such as how suspects refer to victims during denial instances, are also briefly considered, though not with the same systematic approach we conducted for the analysis of prosody and gestures.

Before looking at denial through the lens of multimodality, it is necessary to situate it within the broader context of negation. Hummer et al.'s (1993) study indicates that, from a developmental point of view, denial emerges later than other functions of negation, as it needs the simultaneous representation of two mental models: one reflecting the true state of the world and one reflecting its false counterpart. Equally significant is the study by Ripley (2020) that asserts that denying a certain claim involves performing an act that introduces new information, namely that the claim is ruled out. More broadly, van der Sandt (1991) defines denial as a means of objecting to utterances produced by previous speakers. In this paper, we use the term 'denial' to refer to a speech act encompassing verbal and/or nonverbal elements employed by a speaker to object to or correct the form, content, presuppositions, and implicatures of an utterance (Combei 2023).

This operationalization of denial allows us to examine how denial is encoded multimodally—through prosody and bodily conduct—and how this has been investigated in the literature. First of all, an important contribution to the study of multimodal denial is Harrison's (2018) monograph which argues that negation has clear grammatical and gestural manifestations and that there are regularities between the two elements in human communication. On a similar note, Bressem and Müller's (2017) study on multimodal patterns of negation indicates that recurrent gestures display a fixed form–meaning pairing. It has also been mentioned that multimodality can influence the speech act of denial and their associated belief statuses (Combei 2023). Moreover, the review by Prieto and Espinal (2020) indicates that denial is expressed through various prosodic and gestural features across natural languages, mentioning, in particular, the use of high tones in tonal languages and pitch accentual prominence in intonational languages.

Equally interesting are the studies that explore denial as a deception mechanism from a multimodal perspective, including more recent attempts to automatically detect it. One of the first large-scale multimodal studies on deception is the work by Buller and Aune (1987). They investigate how deceivers manage nonverbal cues to convey nonimmediacy and create a positive image, while simultaneously revealing signs of arousal and negative affect. Buller and Aune's (1987) research, involving 130 participants, claims that deceivers display nonimmediacy and arousal but fail to project a positive image. Additionally, the study indicates that deception cues are influenced by relational history and exhibit significant variability over time. Deceivers also appear to actively regulate their nonverbal behavior, attempting to suppress signs of arousal and negative affect.

A study by Vrij et al. (1996) explores how liars are often unaware that they reduce their movements during deception. Their research aims to determine how deceivers might respond if informed about this rigidity and how factors like tension, behavioral control, and cognitive effort relate to deception. In their experiment, subjects participated in two interviews: one truthful and one deceptive. In the information-present condition, participants were informed beforehand that deception typically involves decreased movement, while the information-absent condition provided no such insight. The findings show that, despite participants believing they increase their movements while lying, they actually exhibit a decrease. Interestingly, informing deceivers about deceptive behavior has no impact on their movements. The authors claim that the decrease is linked to efforts by deceivers to control their behavior and cognitive load, rather than the tension they feel.

Moving to NLP approaches, Soldner et al. (2019) note that deception frequently occurs in everyday conversations, yet conversational dialogues remain underexplored in the field of automatic deception detection. To fill this gap, their paper focuses on detecting multimodal deceptive cues in conversational settings. They introduce a multimodal dataset featuring deceptive conversations from the Box of Lies game on The Tonight Show Starring Jimmy Fallon, where participants attempt to discern whether their opponent's object descriptions are truthful. The authors annotate various multimodal communication behaviors, including facial expressions and linguistic cues, and derive several features from these annotations. Initial classification experiments yield promising results, significantly outperforming both random and human baselines, with an accuracy of up to 69% in differentiating between deceptive and truthful behaviors.

Similarly, Jaiswal et al. (2016) present a data-driven approach for automatic deception detection using audio–video data from real-life trials in legal contexts, focusing, among other things, on visual and verbal cues of denial. They employ OpenFace for facial action unit recognition to analyze witnesses' facial movements during questioning, and OpenSmile to study acoustic patterns. Additionally, the authors conduct a lexical analysis of the spoken words, focusing on pauses and breaks, and feed this data into a Support Vector Machine for deception prediction. They also explore a method that fuses visual and lexical cues through string-based matching. While human judgment accuracy ranged from 53% to 60%, their automated system achieved an average accuracy of 78.95%, with higher accuracy in truth videos (81.10%) than in deceptive ones (76.80%).

As the brief literature review above suggests, previous research has demonstrated that gestures play a significant role in shaping and emphasizing denials, functioning as complementary elements to verbal negation (Harrison 2009). To sum up the overview presented in this section, the multimodal characteristics associated with denials include, among others, head shaking, finger shaking, and palm-down hand gestures (Kendon 2002, 2004; Harrison 2010).

Building upon the research outlined above, this paper seeks to examine the multimodality of denial exhibited in English-language discourse within legal settings in the United States, with a specific focus on individuals accused of femicide (and eventually found guilty). We expect to identify distinctive and systematic patterns of prosodic and gestural features that characterize denials in these specific contexts. The findings of this exploratory analysis may contribute not only to improving our understanding of denial as a linguistic phenomenon, but also to uncovering how it is conveyed through a combination of verbal and nonverbal cues in legal contexts.

The rest of this paper is structured as follows: Section 2 presents the aims, motivations, and scope of the study; Section 3 explains the corpus and methodology; Section 4 outlines the results; and Section 5 provides a discussion of the findings, addresses the limitations of this study, and offers concluding remarks and future directions for our work.

2. Motivations and Aims

This paper presents a qualitative study that is part of a broader research endeavor exploring the multimodal dimension of denial within the legal sphere across the United States. A portion of this larger project, focusing on different data and excluding prosodic analysis, has already been published in Combei (2023). To validate and build upon the findings of the previous study, the present work examines gestural as well as prosodic discursive strategies used by femicide suspects to deny their involvement in crimes during post-crime interactions, such as police interrogations and cross-examinations in courtroom proceedings. The analysis of the suspects' discourse may, in fact, uncover the complex ways in which gendered violence is implicated in denial. This section will explain the rationale of examining the linguistic phenomenon of denial in this specific context, the importance of adopting a multimodal approach in this investigation, and what we aim to achieve with this study.

We concentrate on a specific legal context in which denials of involvement in femicide are uttered, namely situations where the suspect is acquainted with the victim. For the purposes of this paper, femicide is understood as "the killing of women and girls because of their gender" (United Nations 2013, p. 2), as was defined on the International Day for the Elimination of Violence against Women and the Vienna Declaration on Femicide. It should be stressed that femicide differs from general homicide as it is characterized by a disproportionate prevalence of intimate partner violence, familial abuse, and power imbalances (e.g., at home, at work) between victims and perpetrators.

This research centers on the discourse of suspects of femicide precisely because of their close relationship with the victims. We focus on suspects that know the victim well because they may deploy denial strategies that reflect the complex nature of their relationship with the victim (e.g., not admitting or trivializing the severity of the crime, deflecting responsibility, and shifting blame). More generally, analyzing the discourses of this kind of suspect may enhance our understanding of the dynamics of gendered violence and the ways in which such crimes are contested or minimized.

As mentioned above, our study adopts a multimodal perspective on denial, an aspect typically overlooked in forensic linguistics. The term 'multimodality' is used here in accordance with its understanding within the field of conversation analysis and following Mondada's (2016, p. 338) definition as "the various resources mobilized by participants for organizing their action—such as gesture, gaze, facial expressions, body postures, body movements, and also prosody, lexis, and grammar".

As Wang (2024, p. 163) notes, research on legal discourse from a multimodal perspective remains limited, and while gesture studies are advancing in theory and methodology, empirical research in forensic linguistics is still scarce, especially in the area of examining stance in legal discourse through gestures. Some notable exceptions that consider multimodality in analyzing discourse within legal contexts are the studies by Gregory Matoesian. For example, Matoesian and Gilbert (2016) illustrate the importance of multimodal and material actions that accompany speech, showing how attorneys use hand movements, physical objects, and verbal communication to emphasize key pieces of evidence for the jury. The authors also provide a theoretical framework explaining how beat gestures and material objects align with speech to enhance rhythm and highlight points of evidential significance, while also evoking semantic imagery.

The scarcity of multimodal research on legal language is likely attributable to the complexity and time-consuming nature of such analyses, which add to the challenges inherent in investigating legal discourse and content in general. In particular, multimodal analysis of spoken legal language requires the transcription and annotation of a wide range of features, including overlaps, pauses, hesitations, and bodily conduct. In addition, each of these features must be categorized into various classes, each comprising multiple levels (see Section 3 for an example).

Even though we acknowledge the challenges inherent in multimodal analysis, we believe that a close examination of nonverbal features offers a more comprehensive understanding of denial within legal interactions, such as those between suspects and law enforcement. In this regard, we follow Matoesian (2010, p. 541), who asserts that verbal and nonverbal elements function as "co-expressive semiotic partners—as multimodal resources—in utterance construction and the production of meaning". Indeed, the multimodal analysis of discourses produced within legal settings may be useful to better outline the suspects' profiles. With this in mind, our study investigates denial, aiming to describe how suspects negotiate credibility through multimodal resources as well as verbal strategies, before and after a confession or indictment. To this end, the following research question guides our exploratory research: How do suspects of femicide deny allegations?

3. Data and Methods

3.1. Materials

Due to the inherent sensitivity of the data, forensic linguistics corpus collection, storage, access, and distribution are often restricted by privacy legislation (Larner 2019). The ease of access to police recording of custodial interrogations or court data varies from country to country, contingent on the stringency of the pertinent data protection laws. In general, however, building and storing a corpus of forensic data is challenging. For instance, the jurisdictions of Italy and Great Britain impose strict limitations on the accessibility of this type of data (Petyko et al. 2022). As regards the United States, the issue appears to some extent less complex, even though the audio–video recording process and data availability may vary in scope and by state (Bang et al. 2018). A large quantity of forensic multimedia data, such as police interviews, interrogations, cross-examinations, and trials, useful for linguistic analysis can be accessed via online platforms like YouTube. For these reasons it was decided to work with a corpus of multimedia data from North America, in English, and freely retrievable from online sources. The intent was to be able to retrieve an easily accessible dataset that would allow for a focused study of the gestures and prosody of denial.

The entire corpus comprises ten videos sourced from websites and open-access YouTube channels, including *Fifth Estate, Red Circle Interrogations and Confessions, Law & Crime Trial Network*, and *Macon Telegraph Archive*[1]. Five North American suspects, aged 20–44, and accused of femicide are portrayed in the videos; all of them were eventually deemed guilty and convicted. In each instance, the perpetrator was either a close family member or had a close and/or intimate relationship with the victim (husband, boyfriend, or son-in-law). In addition to the accusation of femicide, all the suspects denied the charges on several occasions, some even during and after the trial, appealing the jury verdicts. Four of the suspects were recorded during police interviews. In one case, a suspect was recorded during cross-examination while his trial was in progress.

Initially, the decision to examine denial in two distinct legal contexts (police interviews and cross-examinations in courtroom proceedings) was driven by the goal of conducting a comparative analysis. This comparison was intended to explore how denial functions under different questioning situations. However, as the study progressed, we encountered significant challenges in gathering data from courtroom proceedings, which are scarce, or are not available as high-quality recordings. Given the exploratory nature of our study and its qualitative focus, we adapted our approach. Despite the imbalance in the corpus, we chose to retain the available cross-examination data, recognizing their value in contributing to our understanding of the denial phenomenon, even with a smaller sample size.

The corpus comprises a total of 10,655 tokens, corresponding to a duration of over thirteen hours of audio–video material. The duration and number of tokens in each video were determined by data availability and are, thus, independent of the research design and methodologies implemented. The audio quality of the videos is satisfactory, generally allowing automatic speech processing and analysis of the data. In terms of image quality, some of the data are less satisfactory, and this was reflected in some results (see Section 4). Even if all the videos were recorded in color, in some cases the image resolution was insufficient for the analysis of certain parameters, such as the subtle and swift movement of the eyes and eyebrows. Moreover, although the videos are publicly accessible, all identifying information, including names, sensitive details, and geographic references, were redacted, anonymized, or renamed.

3.2. Methods

The corpus data were used to pursue the examination of gestural manifestations of denial and the analysis of prosody associated with it, before and after the incrimination or admission of guilt. Some aspects related to the pragmatics of referencing the victim and the crime were annotated as comments. The data were processed in accordance with these research directions, so the implementation of distinct procedures was needed.

These steps are summarized below, and each of them is discussed in greater detail in the following paragraphs.

To obtain audio data useful for the analysis of prosodic cues, the .mp4 video files were converted to .wav files using VLC Media Player and Audacity[2]. The resulting audio files were divided into approximately 10-min samples to facilitate the forced alignment process, the .TextGrid creation, and the automatic annotation of pauses through an Automatic Speech Recognition (ASR) pipeline, provided by WebMAUS Services (Kisler et al. 2017).

The pipeline outputted ninety-nine .TextGrid files corresponding to each .wav file considered. Subsequently, a Python script was employed to automatically identify and extract the number and the duration of each pause. The Python script produced Excel spreadsheets, which were used to store the values related to the pauses. The manually extracted information from the .TextGrid files using Praat (Boersma 2001) regarded pitch and intensity. Manual pitch and intensity analysis was preferred in this case, due to the inherent error susceptibility of automated approaches, particularly when considering the quality of the data at hand. The output of the ASR allowed us to use the automatic transcription of the speech as a base for examining the verbal dimension of denial.

As concerns gestural resources, the .mp4 files were processed directly, having been previously annotated with ELAN software[3]. An annotation scheme was designed and implemented using ELAN, with multiple tiers allocated to distinct components of gestural manifestation. Each audio–video track was the object of complex annotation and analysis, with the focus on the conversational turns of the suspect (the process is detailed below).

3.3. Gestures: ELAN and the Annotation Scheme

This study employed ELAN for gestural annotation. A custom annotation scheme was developed to categorize bodily conduct across multiple tiers. Each tier corresponded to a distinct, predefined element created with the controlled vocabulary feature on ELAN. The MIT Boston Speech Communication Group's 'Gesture Coding Manual'[4] was chosen as the annotation scheme for hand gestures. The other features were annotated using the annotation scheme detailed in Combei (2023). We also considered the Linguistic Annotation System for Gesture (LASG), proposed by Bressem et al. (2013) for the annotation of our data. Although well-structured and articulated, we decided not to adopt this annotation scheme because it was too refined, and it took into account some linguistic parameters that were outside the scope of this research (such as syntactic or semantic aspects). At the same time, to the best of our knowledge, LASG lacked annotation patterns for other bodily parameters considered in our study (e.g., head movement, eyebrows, etc.). However, we acknowledge the fact that it would be useful to use LASG for a different, more complex gestural annotation, both to verify the goodness of the scheme we adopted and to explore parameters of multimodality that could not be included in this research. Since this qualitative study relied on a single annotator, future work should involve multiple annotators to measure inter-annotator agreement.

For the purposes of this research, any movement, shape, or orientation expressed by the suspects when uttering a denial act was considered to be a relevant gesture. The types of gestures of interest were restricted to those of the hands (especially their shape and positioning), the head and its movements, the direction of the gaze, the micro-movements of the eyebrows (when analyzable), and the posture during interrogations with police officers or cross-examinations in court. The annotation scheme was designed with six main tiers, each of which was associated with a specific gesture parameter:

1. 'Suspect HG' (hand gesture): describes the shape of the suspect's hand gesture while he is uttering the speech act of denial;
2. 'Suspect Gaze': describes the suspect's gaze direction while he is uttering the speech act of denial;
3. 'Suspect Eyebrows': describes the suspect's movement of the eyebrows while he is uttering the speech act of denial;

4. 'Suspect Head': describes the movement of the suspect's head while he is uttering the speech act of denial;
5. 'Suspect Posture': describes the suspect's body position while uttering the speech act of denial;
6. 'Handedness' (dependent tier of the parent tier 'Suspect HG'): specifies whether one hand or both hands were used to execute the annotated gesture.

Diverging from the annotation scheme developed by Combei (2023), we did not include the 'legs' feature. Due to current resource constraints, we focused our efforts on the upper body and hand movements, ensuring a more in-depth examination of these areas for our research. In fact, we updated the ELAN controlled vocabulary for all of the intra- and inter-suspect recurring movements and positions that were not portrayed in the 'Gesture Coding Manual' (such as 'arms crossing', 'counting', 'measurement', 'pinch').

Furthermore, a tier named 'Suspect' was added for each video to collect the verbal transcript of the suspects' speech (statements). This was used for transcribing their verbal expressions of denial. A tier for comments was also included on the annotation, which was used to highlight relevant elements or findings that went beyond the established labels and annotation scheme. Observations regarding the pragmatics of the suspects' discourses (in particular the way victims and crimes were referenced by the suspects) were also indicated in the comments tier. In order to ensure consistency and facilitate comparison between suspects, the same annotation scheme was used for all videos.

3.4. Prosody

Regarding the prosody of denial, Praat was employed as a tool for speech processing and analysis. Praat functionalities for pitch and intensity analysis were exploited to extract statistic descriptors related to prosodic parameters inherent to the episodes of denial uttered by the suspects. First of all, the intensity was normalized across all videos. Then, minimum, maximum, and average values of pitch and intensity were manually extracted for each instance of denial. To extract these values, we defined the boundaries of each 'denial' instance based on the discursive unit of the suspect. In particular, we considered the discursive unit to be the utterance in which the denial—whether verbal and/or gestural—occurred, extending up to the next pause in the interaction. This approach guaranteed that each denial was analyzed within its immediate context, capturing the correct communicative intent of the suspect.

In terms of speech processing and annotation, Praat was also used to control the pipeline output and check the automatic annotation of pauses. The algorithm's accuracy in identifying the start and end of each pause was evaluated qualitatively and manual intervention was used to correct segment boundaries when necessary. Two primary categories of errors were identified. In the first case, the algorithm failed to accurately identify the onset and conclusion of spoken sequences, resulting in the misclassification of longer segments as pauses. In the second case, the error was more nuanced, involving the inclusion of vowels within the pause segment because the phonation was not correctly captured. All these issues were corrected manually.

4. Results

The research findings will be organized as follows: Section 4.1. will provide a general overview of the analysis with some information regarding the multimodal annotation, the pauses, and some pragmatic observations. Then, Sections 4.2 and 4.3 will be dedicated to gestural and prosodic analysis, respectively.

4.1. General Overview

Table 1 provides a summary of some general results derived from both automatic (i.e., pauses) and manual (i.e., verbal denial) feature annotation. The number of pauses reported for each video depends on the length of the file. The count includes pauses of all types: from those occurring within the same conversational turn to those occurring between

the conversational turns of the suspect and the police officer/lawyer/judge. Thus, we considered both 'pre'-response pauses occurring before the suspect's reply to the official's question and 'post'-response pauses that occur while awaiting the next question or the completion of the suspect's response. In the analyses presented in Section 4.3, we only considered pauses occurring before the suspect's response to questions posed by police officers and lawyers.

Table 1. General Results Statistics—Absolute Frequencies.

Suspect	Videos	Pauses	General Denials	Femicide Denials
A.B.	Video 1	3211	105	70
	Video 2	2369	113	69
I.J.	Video 1	786	106	75
K.L.	Video 1	689	112	75
M.N.	Video 1	896	55	49
O.P.	Video 1	2619	201	180
	Video 2	14,109	235	202
	Video 3	10,526	177	166
	Video 4	1621	88	68
	Video 5	6477	183	162
Total		46,710	1375	1116

Regarding the manual annotation, the fourth and the fifth columns are dedicated to general denials and femicide denials, respectively. This distinction was introduced to account for two-fold manually performed data processing. First, all denial cases encountered during video listening and viewing were annotated, regardless of their degree of relation to the events closely connected to femicide episodes. Subsequently, a manual verification was conducted on these annotations to identify denials expressed by the suspects specifically regarding accusations of committed murder, fictitious statements about the murder weapon, innocence in the matter, concealment of bodies, etc. The column labeled as 'general denials' was included in Table 1 for the purpose of comparison with the column 'femicide denials'. The latter regards the number of denials associated with falsehoods identified in police interviews. This is because isolated denials strictly related to femicides, reported in the fifth column, all turn out to be fictitious denials, intended to distort the reality and avoid a guilty verdict.

'Femicide denials' were identified among the 'general denials' using the following categories as selection criteria: denials related to the timeline of events (for all the events related to the day of the murder itself), specific denials related to the murder weapon (e.g., gun, knife), and denials related to the harm done to the victim (e.g., physical assaults, body concealment). This differentiation between 'femicide denials' and 'general denials' has allowed us to distinguish more clearly between the general use of denial in the forensic context (e.g., the suspect's response "No" to the officer's question "Would you like a glass of water?") and the use of denial for aspects strictly related to the femicides. Below are examples for each identified category of 'femicide denials' to provide insight into the observed data and how it is classified.

1. Timeline of events
    a. Lawyer: Were you in the office when the woman was killed?
       A.B.: No, I wasn't in the office.
    b. Police officer: So why would you call her if you were in the same house. From ten o'clock on. We are not making it up.
       I.J.: No, I'm just saying I'm not recalling this you are talking about.

2. Murder weapon
    a. Police officer: Do you have a gun?
    K.L.: I've never touched a gun before.
    b. Police officer: Did you have other experience where you just wake up and you don't know what happened? Like 'I just woke up and here I am, there was a gun and there was a knife, and drugs and I don't know what was going on', you know, and I understand that.
    O.P.: I've never touched these knives. These knives they were just there. I've never—I've never touched them.
3. Harmed victim
    a. Police officer: You know man, this is stuff we need to know to figure out what's going on.
    M.N.: I didn't try to, I didn't want to, I didn't mean to at all!
    b. Police officer: So what really happened that day?
    O.P.: I didn't do anything. I didn't hurt anybody!

An exploratory review of the transcribed and annotated data revealed some interesting instances of pragmatic choices that align with findings in the work of Combei (2023), providing validation of both studies. In particular, there is an almost complete absence of direct references to the names of the victims in the discourses of the suspects. An educated guess for this vagueness is that the suspects strategically avoid naming the victim to deflect attention and minimize their emotional engagement as well as the consequences of the crime.

The victims' names occur in only three cases throughout the entire 13-h corpus. In example (4), the victim's name is uttered as a response to a direct and explicit question from the police officer, in example (5), the name appears only as an appellation used in reported speech, and finally in example (6), the name is uttered as a violent way to distance oneself from the accusations made by the police officer. In all other instances where the suspects mentioned the victim, they used anaphoric expressions, typically referring to the victims with third-person singular pronouns (i.e., she or her).

4. Police officer: And what's your wife's name?
    K.L.: Claire.
5. Police officer: And what did you do next?
    O.P.: We said like "Have you talked to Jo?" I was like "No, have you talked to Jo?"
6. Police officer: O. you are under arrest for murder right now. The murder of Johanna.
    O.P.: I didn't murder Johanna! I don't.

On the same note, another interesting pragmatic aspect is the lack of direct references to the act and the result of killing in the suspects' discourses. In fact, in the annotated speech (Suspects tier), the word 'murder' appears only once (uttered by one suspect), while terms like 'death', 'to kill' (0), and 'dead' are entirely absent from the dataset. Instead, we frequently find generic pronouns, names, verbs, or other anaphoric expressions used to refer to the crimes and their consequences. For instance, terms such as 'it' (126), 'that' (117), and 'anything' (75) occur frequently, as expected, especially following explicit references to the crime made by police officers. In this case, the vagueness could also be interpreted as both a mitigation strategy (i.e., it lessens the weight of the crime) and a detachment strategy from the victim.

4.2. Gestural Analysis

Regarding the gestural manifestation of denial, as detailed above, we annotated all videos based on posture, hand gestures, gaze, eyebrows, and head movements. The output of the annotation process allowed us to extract the most frequent features for various bodily characteristics,[5] namely 'front' for head position, 'sitting erect' for posture, 'towards other speakers' for gaze, 'open' for hand gestures, and 'both' for handedness. It should be mentioned that the frequency of occurrence of each type of feature is influenced by the

different sizes of the five subcorpora; in particular, the disproportionate amount of data annotated for the O.P. suspect skews the final count of each entry. To address this issue and account for the specific distribution of the gestural features, information related to the subcorpus of each suspect considered is reported in Table A1 in Appendix A.

Given the frequency of gestures and the acknowledged imbalance in our data, the results, though interesting, should be interpreted with caution. That being said, our findings largely point to a prototypical multimodal expression of denial, characterized predominantly by a simple gestural apparatus involving the head either in a frontal position or lowered downwards (presumably to obscure the gaze). Head shaking occurs parallel to the downward head movement. This gesture occurs multiple times even autonomously, serving as a paraverbal signal of denial without any verbal expression. Following the head gestures, the gaze is engaged, primarily in the configuration 'towards other speaker' (mostly in concordance with the head in a frontal position) or avoiding, opposing the interlocutor's gaze by maintaining a closed-eye configuration throughout the denial. Additionally, similar to the 'avoiding disposition' that characterizes closed eyes, there is the 'down' configuration, predominantly occurring with a 'frontal' head position.

The significant number of cases where annotation of eyebrow-related traits was not feasible (due to the video recording quality or the angles) makes this parameter challenging to assess. This highlights the importance of using high-quality multimedia material in multimodality studies. The most suitable situation for annotating eyebrows was found in the video of suspect A.B., filmed in the courtroom, where the high-quality close-up footage allowed precise observation of facial expressions, shapes, and movements. Despite technical issues, it is interesting to note that, in the cases where these features could be analyzed, denial did not manifest through eyebrow movements support ('relaxed' featured 179 occurrences). Nevertheless, paradoxically, within this dataset the feature 'frowning' often occurs, denoting a very specific movement typically associated with negative emotions.

The most common posture associated with denial is the suspect seated, often upright (erect), and facing the interviewer. However, this posture is prevalent across the entire corpus and is not exclusive to denial scenarios. An interesting pattern in the corpus involves suspect O.P., who frequently adopts a forward-leaning posture, particularly after the confession, conducting much of the interview hunched over. Since this posture is especially traceable in the second phase of the interviews, one plausible interpretation is that the suspect may feel increasingly pressured.

The situation concerning hand gestures is more complex. To begin with, we will focus on the less complex findings. Our observations indicate that the majority of manual gestures associated with denials are predominantly performed with both hands simultaneously. It would be interesting to investigate whether this aspect is specific to the multimodality of denials (and the forensic contexts) or if what is observed is a general trend in gestural expression. Next, moving to more complex aspects of hand gestures, we found that the most frequent categories for hand gestures are the forms 'open', 'relaxed', and 'arms crossed'. We have intentionally excluded the 'fist' hand gesture from what we report as prototypical denial gestures, despite their frequency. In fact, the 'fist' label appears predominantly in O.P. videos and is therefore more indicative of individual expression of denial rather than representative of broader patterns of this phenomenon. The 'handcuffed' label is excluded from our analysis because the mere presence of restrained hands does not qualify as a gesture. Handcuffs represent a state of limitation rather than an intentional communicative action. During the final "verdict" we traced an open-hand gesture performed with both hands while keeping the head in a frontal position in relation to the interlocutor. This is accompanied by a gaze mostly directed towards the interlocutor, while seated in an upright position.

As shown in Table A1 in Appendix A the 'not available' category of gestures is very frequent as a result of A.B.'s framing in the video, which predominantly features close-up shots that obscure the visibility of arms and hands. It is important to mention that the 'not available' label applied to all types of gestures (e.g., posture, gaze, etc.) does not represent a real feature. However, we documented instances where gestures were indistinguishable due to video quality or subtle movement limitations in order to capture the impact of visibility constraints on our findings.

An important point to emphasize is the difference between the gestural multimodality of denials as it occurs before the confession or 'turning point'[6] during the police interview and the multimodality of denials as it occurs after these relevant moments. In our dataset there is a marked reduction in both verbal and nonverbal expression in the 'post'-confession phases to varying degrees among all suspects. Below is an example illustrating both the verbal and multi-modal behavior of one suspect at two distinct moments: first, prior to learning about their partner's death resulting from their aggression, and subsequently, following confirmation and subsequent charges of femicide.

As can be inferred from Figure 1, in the 'pre'-phase, the suspect's conversational turn is marked by heightened gestural dynamism, complemented by generally longer and more complex sentences. In Figure 2, however, greater heaviness and stillness is observed in the physicality and gestures of the suspects. In the 'post'-phase the curtailment in verbal expression is total, as nothing is uttered verbally; head shaking is the only element through which the suspect conveys his denial. It is interesting that the open hand shape is clearly visible, suggesting, in this case, an attitude of non-acceptance of the facts.

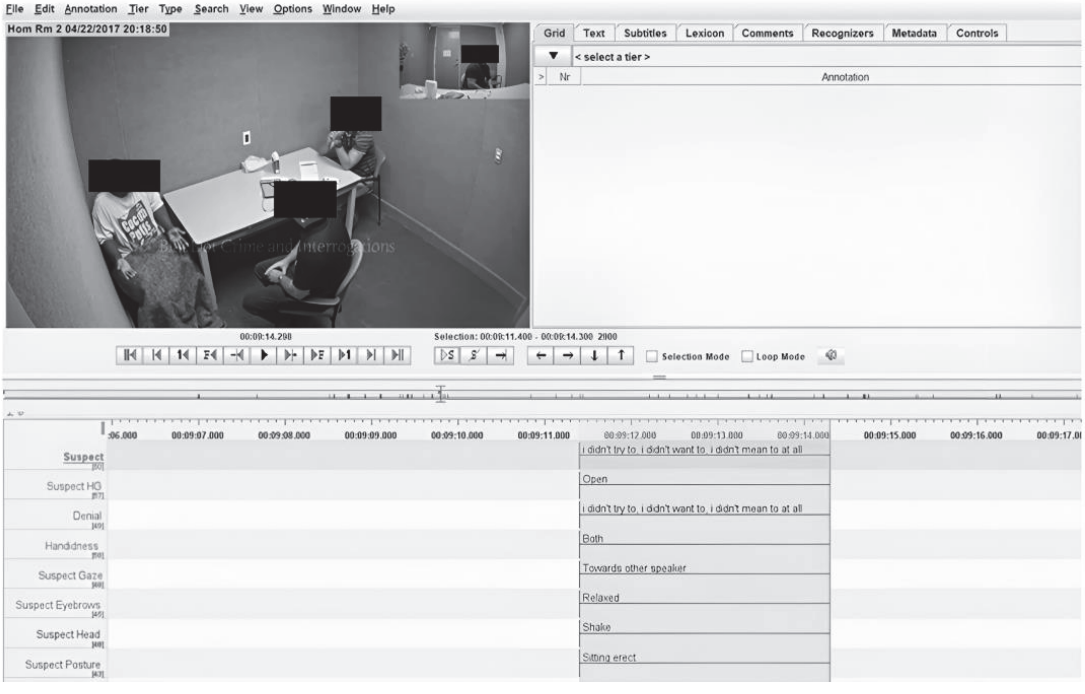


Figure 1. M.N. Multimodal Denial ('pre')[7].

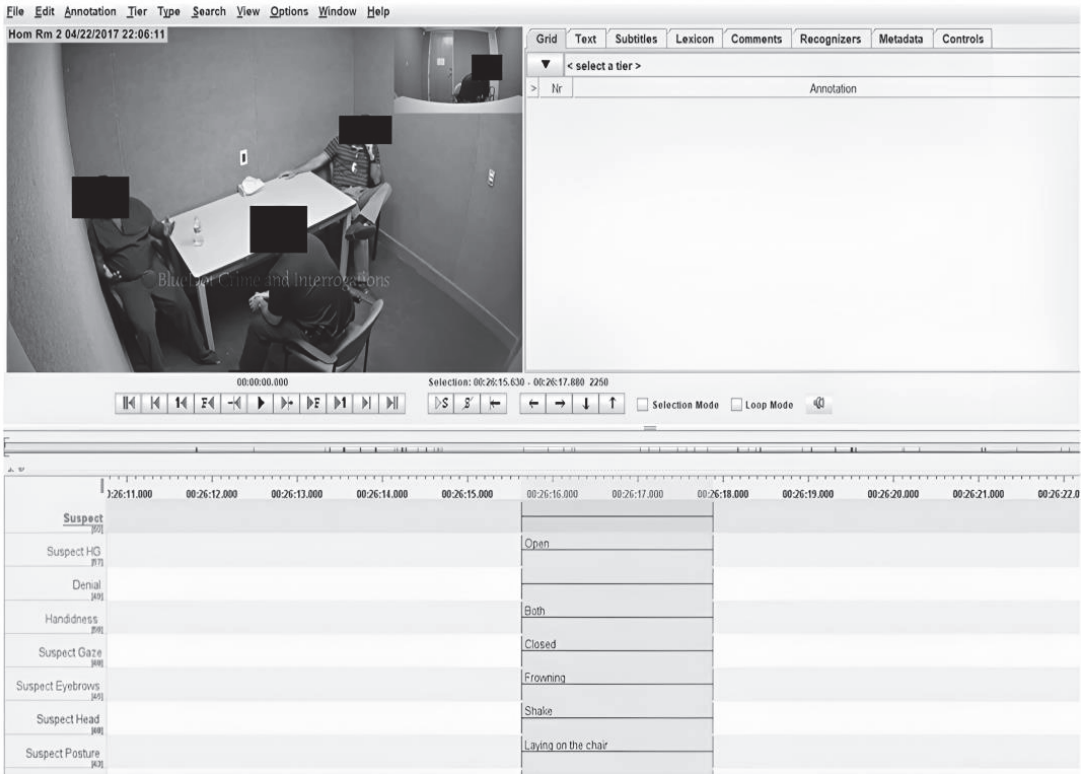


Figure 2. M.N. Multimodal Denial ('post').

4.3. Prosodic Analysis

The prosodic analysis involved extracting the values of minimum, mean, and maximum pitch and intensity for each speech act of denial. This manual extraction was complemented by automatic extraction of pauses. Given the significantly large amount of annotated data, we decided to work with the extracted values of ten sample denials selected for each suspect (the denials with their pitch and intensity values were first stored in an Excel spreadsheet and then randomly selected to avoid bias). All these ten samples were selected from the subset of femicide denials. For the denials selected for each suspect, a further internal subdivision of the total collected denials was made. Denials produced before the confession or turning point during police interviews were distinguished from denials produced after these moments. Of the ten denials selected for each suspect, five were randomly chosen from the 'pre-confession' denials, while the remaining five were selected from those produced by the suspect during 'post-confession'. This choice is motivated by the interest in the variation of pitch, intensity, and the number of pauses between, before, and after the confirmation of the accusations or the suspect's admission of guilt. This is aimed at observing a possible systematic difference in the parameters between before and after instances, motivated by emotional and circumstantial reasons stemming from the exposure of lies and/or the formal accusation of femicide. It was assumed that there could be variation due to the strong emotional impact that being caught lying and/or being accused of murder entails, and that this could be found in all cases under consideration. Even if this possibility is acknowledged, its quantification falls outside the aims of this study.

Table 2 shows the 'pre' and 'post' averages for pitch and intensity for each suspect. The results appear to provide some responses to the research question. In particular, variation is

observed, within the constraints of an exploratory qualitative study, regarding the intensity and the pitch of denial expression. In two cases, it seems to be more contained regarding pitch (K.L. and O.P.). What should be noted is that this snapshot of observations does not seem to indicate a steady direction of variation for the parameters considered, particularly concerning pitch. For intensity, there is a tendency towards a decrease in decibels following the confession or turning point (and during the cross-examination phase) in four out of the five suspects. O.P. contradicts this trend with a significant deviation and an increase in intensity after the turning point of the interview. Regarding pitch, however, there is a tendency towards a decrease following the confession or during the cross-examination phase in the cases of A.B. and I.J., while for the remaining three, there is an increase in F0 within the same circumstances. It is certainly important to mention, however, that the only case showing a significant increase in pitch is M.N., while K.L. and O.P. present a more subtle variation in which the increase may be more due to randomness or idiosyncrasy.

Table 2. Mean of Pitch and Intensity.

Suspect	Average Pitch ('Pre')	Average Pitch ('Post')	Average Intensity ('Pre')	Average Intensity ('Post')
A.B.	123.39 Hz	119.01 Hz	69.11 dB	65.05 dB
I.J.	160.10 Hz	151.81 Hz	62.46 dB	59.57 dB
K.L.	220.70 Hz	221.53 Hz	62.69 dB	59.32 dB
M.N.	175.02 Hz	208.81 Hz	79.40 dB	62.81 dB
O.P.	174.80 Hz	176.26 Hz	69.66 dB	81.17 dB

Table 3 shows the average length of pauses calculated by 'pre' and 'post' phase, in addition to the analysis of pitch and intensity. Regarding the average length of pauses observed in individual suspects, it appears that there are longer pauses in the post-phase for A.B., I.J., and M.N. Even if K.L. and O.P. do not confirm this trend, the gap between the 'pre' and 'post' phase in these two suspects is smaller than the gap observed in the other three. The 'pre' and 'post' totals reflect the majority trend. Regardless of the length, the number of pauses is almost equivalent in the 'pre' and 'post' phases for all five suspects. However, this does not correspond with the distribution of pauses in the entire dataset. In general, the number of pauses is greater in the 'post' phases of each suspect.

Table 3. Average Length of Pauses.

Suspect	Average Pauses Length ('Pre')	Average Pauses Length ('Post')
A.B.	0.94 s	1.51 s
I.J.	0 s	0.69 s
K.L.	1.23 s	1.04 s
M.N.	0.56 s	3.41 s
O.P.	1.11 s	0.81 s
Total	0.77 s	1.49 s

5. Discussion, Limitations, and Conclusions

Although the exploratory nature of this qualitative study precludes the derivation of systematic generalizations, the results and the interpretation of the data described in Section 4 highlighted the prototypical nature of both generic and feminicide-specific multimodal expressions of denial. Here, we extend the above considerations by adding some comparisons with the relevant literature on gestures, particularly in relation to arms and hands movements.

Concerning arms, we interpret the feature 'arms crossed' as a gesture of closure and separation, typically suggestive of downplaying and avoidance (Gallace et al. 2011). Therefore, we can claim that this element functions as a mechanism to express detachment and diminish the perceived importance of the crime in question, which is rendered as unexpected. Thus, this could indicate an intentional effort to downplay involvement in crime-related events.

In Figures 3 and 4 we see two examples of the 'arms crossed' gesture, which we interpret as a cue of detachment from the crime under discussion. In the first case (Figure 3), the position of the arms co-occurs with a fake statement of desperate and sad astonishment ("I don't know why anybody would do that"), aimed at avoiding possible allegations of involvement. In the second example (Figure 4), the closed position with crossed arms is also used by the suspect to detach himself from the reality of the situation. In this particular instance, the selected image represents a frame within a 'bump and grind' phase of the police interview. During this phase, the suspect assumes and maintains a defensive position, responding to all questions posed by the police officer with a lie. More generally, at the corpus level, this gesture appears to be associated with a defensive stance.

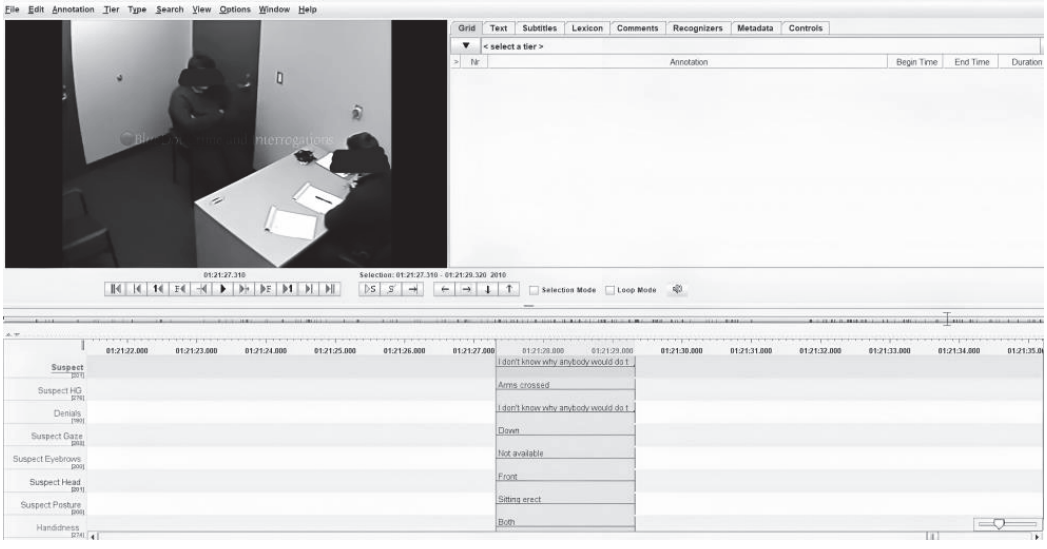


Figure 3. O.P.'s example of denial accompanied by the 'arms crossed' gesture.

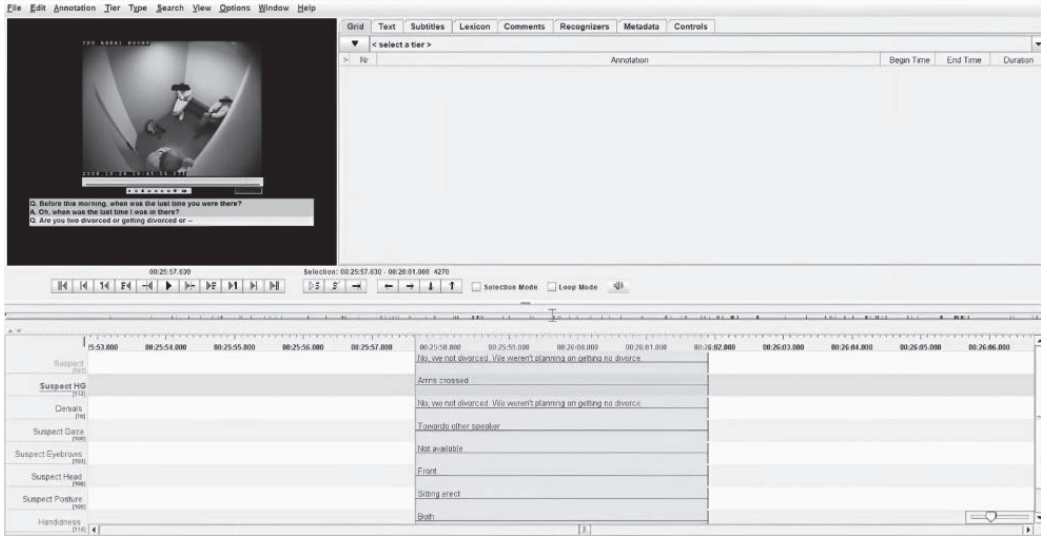


Figure 4. I.J.'s example of denial accompanied by the 'arms crossed' gesture.

Similarly, an 'open' hand gesture, in both configurations (palm up and palm down), falls under Kendon's (2004) classification which analyzes this type of hand shape in relation to manual actions of stop, refusal, denial, or interruption accompanying verbal expression. It is significant from this perspective that this specific hand shape is the most recurring one, not only throughout the dataset of denials but also for denials specifically related to femicides.

Two particularly illustrative examples are provided below. The 'open' hand gesture, employed in both the palm-up and palm-down configurations, was used by the same suspect to explicitly disavow any involvement with the murder weapon. In Figure 5, the suspect's hands, with palms facing upwards, accompany the declaration of innocence concerning the allegation of firearm usage. The use of the hands creates a particular sense of surrender and innocence that follows the sentence, which not only denies the use of a firearm but also its possession ("I've never had a gun").

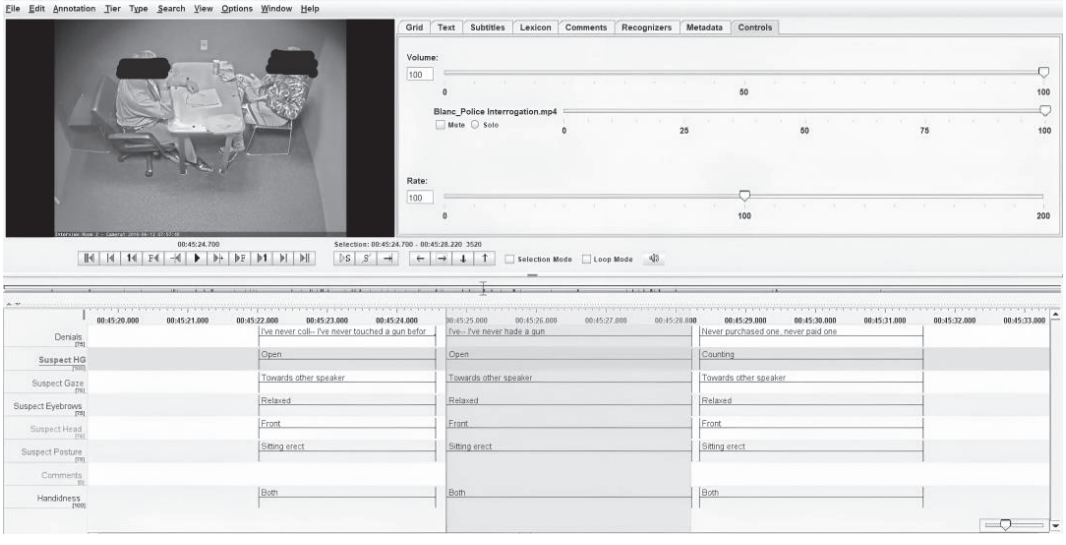


Figure 5. 'Open' hand gesture (palm up).

Figure 6 illustrates the progression of denial. After discrediting the initial assertion regarding contact with a firearm, the police officer asks about the location of the firearms at the crime scene. The suspect then denies having used them, stating, "Please, don't take it the wrong way. I really never used it. Neither one of them." In this instance, the suspect's use of the phrase "Neither one of them" accompanied by the gesture of shaking his open hands with the palm down in front of him, serves to reinforces his denial and strengthens his assertion that he has no connection to the murder weapon.

Then, we observed the 'relaxed' gesture, which to some extent resembles the 'open' form, occurring mostly in conjunction with other gestures. It was indeed the hand shape most often adopted individually by the right or left hand and not in double configurations.

As previously stated, the examples confirm the assumption of the 'relaxed' hand shape within a gesture made with only one of the two hands. In the case of Figure 7, the relaxed right hand is accompanied by a gesture of nervousness (the act of scratching the back or face) that O.P. often performs when he is in a recumbent position in relation to his actions on the day of the crime. In Figure 8, on the other hand, the suspect's right hand, in a relaxed position, is accompanied by the left hand that instead expresses denial in an open position with palm facing upward. The 'open' form is once more employed to convey innocence and detachment from the facts being verbally denied, as illustrated in Figures 5 and 6.

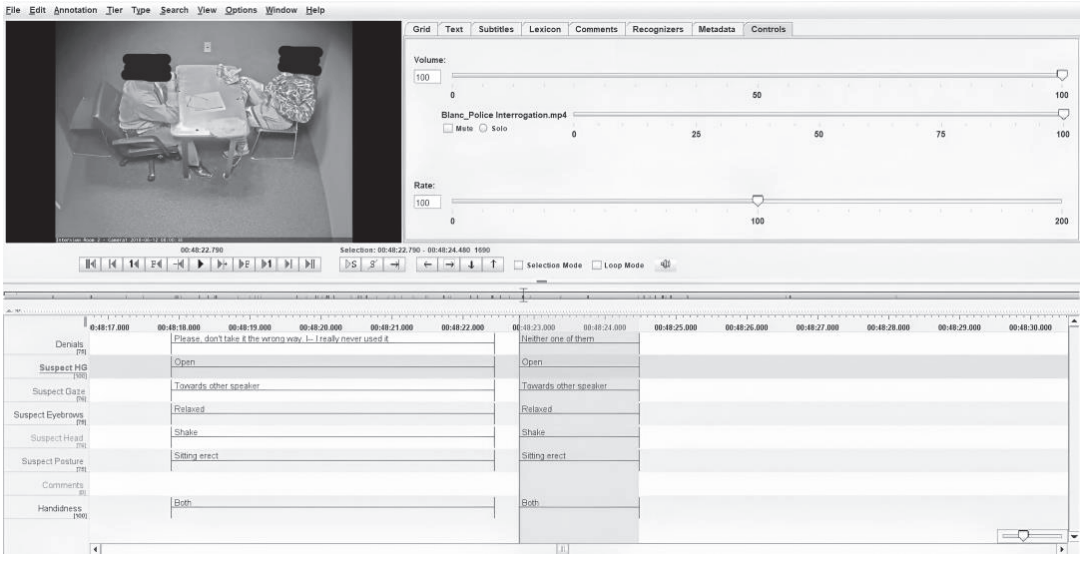


Figure 6. 'Open' hand gesture (palm down).

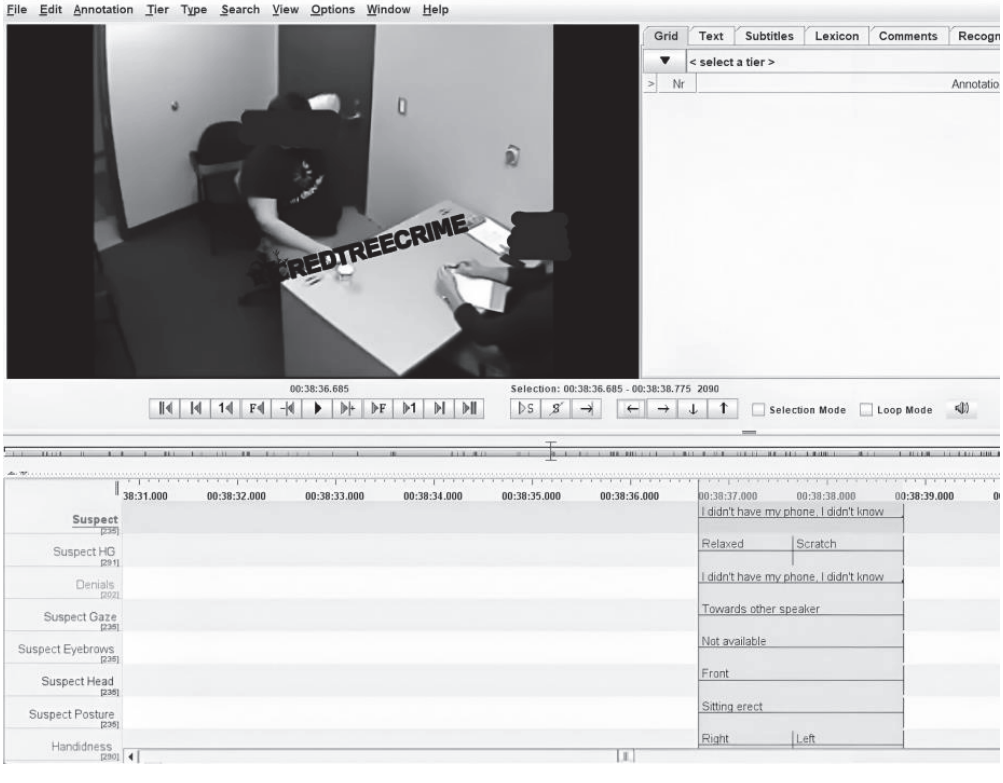


Figure 7. O.P.'s example of a 'relaxed' gesture.

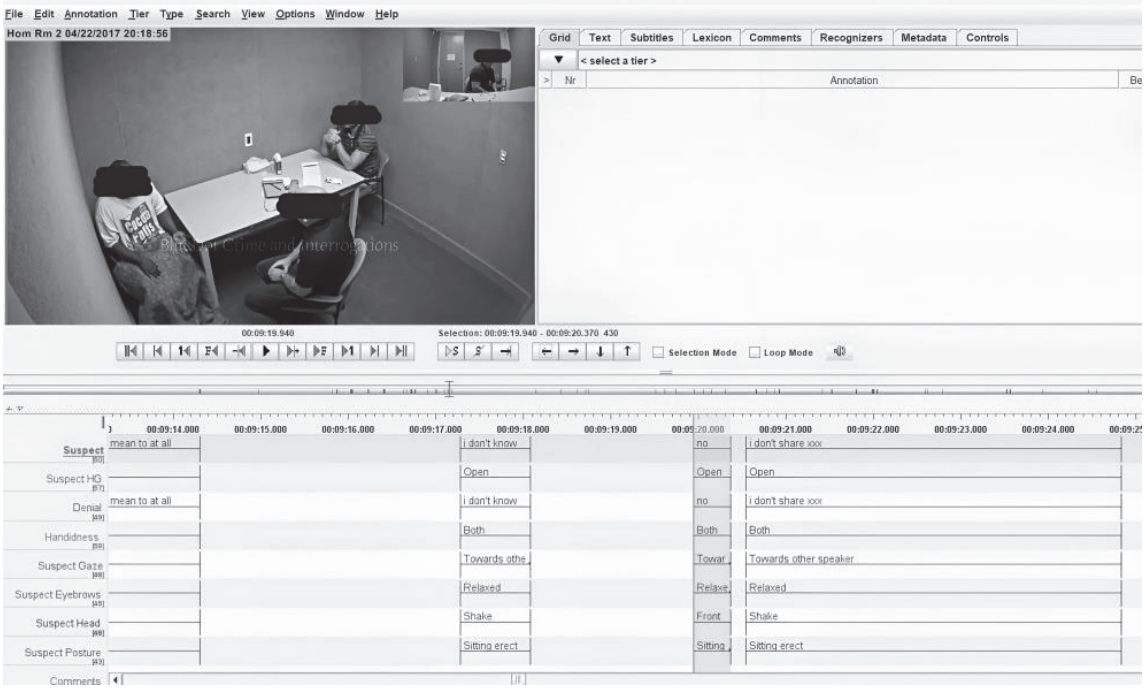


Figure 8. M.N.'s example of a 'relaxed' gesture.

Having presented and discussed our results, it is necessary to acknowledge that this study is preliminary in nature, and it has several limitations, leaving numerous areas for further exploration and refinement. Some of the limitations may, in fact, represent new opportunities for development, which could be implemented in the relatively near future.

First of all, for our findings to be supported quantitatively, the study would require a larger and more balanced corpus (e.g., the same amount of data for police interviews and cross-examinations in courtroom proceedings). At the same time, a greater amount of audio–visual material from suspects of femicide, as well as from people accused of other crimes, is needed to compare the features assessed in this study with those in cases involving different types of criminal suspects. This would provide a better picture of the pragmatic, multimodal, and prosodic behavior of suspects, providing a more representative and generalizable overview of how denial is expressed verbally and/or nonverbally.

From a multimodal perspective, while the annotation scheme used in this study was sufficiently rich and complex, we believe it could be further enhanced by adding a few new parameters. Particularly, it could be useful to add cues regarding the lower body as well as 'shoulder shrugs' as a gestural feature. As reported by previous research, lower body cues (e.g., feet and legs) are less studied compared to the upper body but they are still relevant for the organization of social interaction (Mondada 2014). Equally interesting are the features regarding the shoulders: including them into the annotation of denial could be useful because the action of shoulders shrugging has been reported to signal a detachment attempt since "[they] can work as markers of 'dis-stance' or disengagement, in which case they take on an epistemic-evidential dimension" (Debras 2017, p. 24).

Since our study did not investigate the detailed shape and execution of gestures, future research could benefit from decomposing these gestures into finer traits, configurations, and subtler movements. This would offer a more granular picture of the multimodality of denial and, in more general terms, it would represent a multi-level analysis of gestures, investigating not only the syntactic–semantic or pragmatic aspect but also the "morphological" composition of gestures. At this stage, as outlined in the methodological section, the inclusion of an additional annotator and calculating inter-annotator agreement would be necessary for validation purposes.

As regards prosody and focusing particularly on the fundamental frequency parameter, another possibility to enhance this kind of study is to carry out precise annotation of the intonational contour of denials, to assess the possible presence of denial-specific pitch characteristics. In this regard, Mertens' (2014, 2020) Prosogram and Polytonia tools could be used to automatically obtain a stylization of pitch contour as well as an automatic labelling of pitch movements. Parallelly, following the line drawn by classical works of Beckman and Pierrehumbert (1986) as well as Ladd (2008) on the intonational aspects of the English language (particularly the study of pitch accents and pitch contours based on the autosegmental-metrical theory), specific configurations could be observed, which may also be useful for an analysis of the pragmatic use of intonational features, both in production and perception. Moreover, once the necessary conditions for the retrieval of the aforementioned useful data are satisfied, comparisons could be drawn with studies on intonational contours of denials in other languages, such as Italian and other Romance languages (D'Imperio 2002; Prieto et al. 2005). Additionally, considering both prosodic and pragmatic aspects together, it would be interesting to complete the data on the mean length of utterances, adding speech rate (e.g., in the form of a count of syllables produced per second by each suspect), both overall and separately in the 'pre' and 'post' turning point phases.

Despite the limitations discussed above, the results of our paper outline a recognizable profile for multimodal denial. Recurrent features include a predominant gestural component characterized by head positioning (either neutral or lowered) and head shaking. The head shaking feature is frequently repeated and can serve independently as a nonverbal marker of denial. We also observed that denial is often accompanied by open-hand gestures and a sitting (erect) posture; this posture is frequently observed in conjunction with denial but is also common in non-denial instances throughout the corpus. A certain degree of vagueness in speech patterns was reported as regards the way suspects refer to the victims (i.e., they are not named explicitly). As regards prosody, our findings indicate that expressions of denial frequently involve a reduction in pitch and intensity following a confession or indictment. Finally, the analysis of pauses reveals that a greater number of pauses typically occur after incrimination.

Overall, we believe that this research may contribute to future studies on the multimodality of denial in legal settings, and to the limited literature in forensic linguistics and the broader academic discourse on this topic.

Author Contributions: Conceptualization, E.D. and C.R.C.; Methodology, E.D. and C.R.C.; Software, E.D. and C.R.C.; Validation, C.R.C.; Formal analysis, E.D.; Investigation, E.D. and C.R.C.; Resources, E.D. and C.R.C.; Data curation, E.D.; Writing—original draft, E.D. and C.R.C.; Writing—review & editing, E.D. and C.R.C.; Visualization, E.D.; Supervision, C.R.C.; Project administration, C.R.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The audio–video material used in this study cannot be made available as the videos were retrieved from online sources and are not owned by the authors. The code used for the automatic feature extraction developed for this study is available under request, contacting the corresponding author.

Acknowledgments: This research was conducted while Claudia Roberta Combei was engaged as a researcher under the initiative PON Ricerca e Innovazione 2014–2020—Linea Innovazione (D.M. 1062/2021) and while Elena Didoni was an M.A. student at the University of Pavia, where this work formed the basis of her thesis. The authors would like to thank Lucia Busso and Ilaria Fiorentini, as well as the four anonymous reviewers, for their helpful comments and suggestions on this study.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Table A1. Multimodality Occurrences (Denials Femicides).

Body Part	Annotation	Whole Corpus	I.J.	K.L.	M.N.	A.B.	O.P.
Head	Back	10	0	0	0	0	10
	Front	467	38	41	7	50	331
	Left	72	0	3	2	2	65
	Nod	0	0	0	0	0	0
	Right	50	0	0	2	0	48
	Shake	187	10	32	35	87	23
	Down	235	26	0	0	0	209
	Head bent right	34	2	0	1	0	31
	Head bent left	58	2	0	0	0	56
	Not available	7	2	0	0	0	5
Gaze	Blink	0	0	0	0	0	0
	Down	290	42	3	12	5	228
	Eyerolls	0	0	0	0	0	0
	Front	25	1	0	2	0	22
	Left	68	0	0	0	2	66
	Right	10	0	0	0	0	10
	Towards other speaker	376	35	72	25	130	114
	Up	3	0	1	0	2	0
	Wink	0	0	0	0	0	0
	Closed	381	0	0	4	0	377
	Not available	29	2	0	0	0	27
Eyebrows	Both eyebrows raising	85	0	11	2	63	9
	Frowning	170	0	1	3	11	155
	Left eyebrow raising	7	0	1	0	6	0
	Relaxed	179	0	62	23	34	60
	Right eyebrow raising	23	0	0	0	23	0
	Not available	643	75	0	15	0	553
Posture	Laying on the chair	156	23	1	10	0	122
	Leaning forward	333	18	1	3	22	289
	Moving left	4	0	0	0	0	4
	Moving right	20	1	0	0	1	18
	Retracting back	12	1	0	1	0	10
	Sitting erect	718	52	109	0	182	375
	Swivel chair	4	0	0	0	0	4
	Not available	6	2	0	0	0	4

Table A1. *Cont.*

Body Part	Annotation	Whole Corpus	I.J.	K.L.	M.N.	A.B.	O.P.
Hand Gesture	Angled	4	0	2	0	0	2
	Arms crossed	197	24	0	0	21	152
	Ball	1	0	1	0	0	0
	Clasped	39	7	3	0	14	15
	Counting	4	0	1	0	0	3
	Cross	1	1	0	0	0	0
	Cup	8	1	2	0	0	5
	Deictic	18	2	0	0	1	15
	Fist	108	0	2	8	0	98
	Folded	8	7	0	0	1	0
	Gun	0	0	0	0	0	0
	Hands in pockets	83	26	54	0	0	3
	Hole	0	0	0	0	0	0
	Intertwined	19	0	0	0	1	18
	Jailed	0	0	0	0	0	0
	Knife	11	2	3	0	0	6
	Loose	30	0	0	0	1	29
	Measurement	2	0	2	0	0	0
	Okay	0	0	0	0	0	0
	Open	311	4	17	35	16	239
	Pursued	0	0	0	0	0	0
	Relaxed	151	3	10	13	11	114
	Scratch	79	2	1	0	0	76
	Star	5	0	0	0	0	5
	Steepled	1	0	0	0	0	1
	Tulip	0	0	0	0	0	0
	Two	0	0	0	0	0	0
	Wall	47	0	0	0	0	47
	Finger closed	24	0	0	0	0	24
	Pinch	26	0	0	0	0	26
	Fingers touching	8	0	0	0	0	8
	Drinking	1	0	0	0	0	1
	Handcuffed	116	0	0	0	0	116
	Writing	1	0	0	0	0	1
	Key	1	0	0	0	0	1
	Not available	100	5	0	0	73	22
Handedness	Both	792	70	56	37	56	573
	Left	258	5	22	10	3	218
	Right	276	8	22	10	7	229
	Not available	74	1	0	0	69	4

Notes

1 These channels are available at the following web pages (accessed 18 July 2024): https://www.cbc.ca/news/fifthestate; https://www.youtube.com/@Telegraph247; https://lawandcrime.com/; https://www.youtube.com/@redcircleinterrogationsand5722.

2 The software are available at the following web pages (accessed 18 July 2024): https://www.videolan.org/vlc/index.it.html; https://www.audacityteam.org/.

3 This software is available at this web page (accessed 18 July 2024): https://archive.mpi.nl/tla/elan.

4 The MIT Boston Speech Communication Group's 'Gesture Coding Manual' is available at this web page (accessed 18 July 2024): https://speechcommunicationgroup.mit.edu/gesture/coding-manual.html.

5 With respect to eyebrow features, the 'not available' category was the most frequently assigned due to unsatisfactory image quality or recording angles.

6 By 'turning point' we mean the moment of the police interview when the suspect's alibi is overtly challenged or debunked. At that point, the suspect must face the truth and decide whether to continue with his strategy or confess to the crimes.

7 In accordance with what had been stated in lines 184–186, it was resolved that the faces of suspects depicted in the video would be obscured in deference to the right to privacy, even though the multimedia material is freely accessible online. All figures in the

paper were processed following this methodology. To view the multimedia data in full, interested parties may request the image from either of the paper's authors.

References

Bang, Brandon L., Duane Stanton, Craig Hemmens, and Mary K. Stohr. 2018. Police recording of custodial interrogations: A state-by-state legal inquiry. *International Journal of Police Science & Management* 20: 3–18. [CrossRef]

Beckman, Mary Esther, and Janet Breckenridge Pierrehumbert. 1986. Intonational structure in Japanese and English. *Phonology* 3: 255–309. [CrossRef]

Boersma, Paul. 2001. Praat, a system for doing phonetics by computer. *Glot International* 5: 341–45.

Bressem, Jana, and Cornelia Müller. 2017. The "negative-assessment-construction"—A multimodal pattern based on a recurrent gesture? *Linguistics Vanguard* 3: 1–9. [CrossRef]

Bressem, Jana, Silva H. Ladewig, and Cornelia Müller. 2013. Linguistic annotation system for gestures (lasg). In *Body—Language—Communication: An International Handbook on Multimodality in Human Interaction*. Edited by Cornelia Müller, Alan Cienki, Ellen Fricke, Silva H. Ladewig, David McNeill and Jana Bressem. Berlin: De Gruyter Mouton, pp. 1098–124.

Buller, David B., and R. Kelly Aune. 1987. Nonverbal cues to deception among intimates, friends, and strangers. *Journal of Nonverbal Behavior* 11: 269–90. [CrossRef]

Combei, Claudia Roberta. 2023. The multimodal expression of denial: A case study on femicide suspects. In *Linguistica Forense in Prospettiva Multidisciplinare*. Edited by Sonia Cenceschi and Chiara Meluzzi. Milano: Officinaventuno, pp. 27–44.

Debras, Camille. 2017. The shrug. *Gesture* 16: 1–34. [CrossRef]

D'Imperio, Mariapaola. 2002. Italian intonation: An overview and some questions. *Probus* 14: 37–69. [CrossRef]

Gallace, Alberto, Diana M. E. Torta, G. Lorimer Moseley, and Gian Domenico Iannetti. 2011. The analgesic effect of crossing the arms. *Pain* 152: 1418–23. [CrossRef]

Harrison, Simon. 2009. The expression of negation through grammar and gesture. In *Studies in Language and Cognition*. Edited by Jordan Zlatev, Mats Andrén, Marlene Johansson Falck and Carita Lundmark. Cambridge: Cambridge Scholars Publishing, pp. 421–35.

Harrison, Simon. 2010. Evidence for node and scope of negation in co-verbal gesture. *Gesture* 10: 29–51. [CrossRef]

Harrison, Simon. 2018. *The Impulse to Gesture: Where Language, Minds, and Bodies Intersect*. Cambridge: Cambridge University Press.

Horn, Laurence Robert. 2010. *The Expression of Cognitive Categories*. Berlin: De Gruyter.

Hummer, Peter, Heinz Wimmer, and Gertraud Antes. 1993. On the origins of denial negation. *Journal of Child Language* 20: 607–18. [CrossRef]

Jaiswal, Mimansa, Sairam Tabibu, and Rajiv Bajpai. 2016. The truth and nothing but the truth: Multimodal analysis for deception detection. Paper presented at 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), Barcelona, Spain, December 12–15; pp. 938–43.

Kendon, Adam. 2002. Some uses of the head shake. *Gesture* 2: 147–82. [CrossRef]

Kendon, Adam. 2004. *Gesture: Visible Action as Utterance*. Cambridge: Cambridge University Press.

Kisler, Thomas, Uwe Reichel, and Florian Schiel. 2017. Multilingual processing of speech via web services. *Computer Speech & Language* 45: 326–47. [CrossRef]

Ladd, Dwight Robert. 2008. *Intonational Phonology*. Cambridge: Cambridge University Press.

Larner, Samuel. 2019. Formulaic sequences as a potential marker of deception: A preliminary investigation. In *The Palgrave Handbook of Deceptive Communication*. Edited by Tony Docan-Morgan. London: Palgrave Macmillan, pp. 351–70.

Matoesian, Gregory M. 2010. Multimodal aspects of victim's narrative in direct examination. In *The Routledge Handbook of Forensic Linguistics*. Edited by Malcolm Coulthard and Alison Johnson. New York and London: Routledge, pp. 541–57.

Matoesian, Gregory, and Kristin Gilbert. 2016. Multifunctionality of hand gestures and material conduct during closing argument. *Gesture* 15: 79–114. [CrossRef]

Mertens, Piet. 2014. Polytonia: A system for the automatic transcription of tonal aspects in speech corpora. *Journal of Speech Sciences* 4: 17–57. [CrossRef]

Mertens, Piet. 2020. The Prosogram model for pitch stylization and its applications in intonation transcription. In *Prosodic Theory and Practice*. Edited by Jonathan Barnes and Stefanie Shattuck-Hufnagel. Cambridge, MA: MIT Press, pp. 259–86.

Mondada, Lorenza. 2014. Bodies in action. *Language and Dialogue* 4: 357–403. [CrossRef]

Mondada, Lorenza. 2016. Challenges of multimodality: Language and the body in social interaction. *Journal of Sociolinguistics* 20: 336–66. [CrossRef]

Morris, Bradley J. 2003. Opposites attract: The role of predicate dimensionality in preschool children's processing of negations. *Journal of Child Language* 30: 419–40. [CrossRef]

Petyko, Marton, Lucia Busso, Tim Grant, and Sarah Atkins. 2022. The Aston Forensic Linguistic Databank (FoLD). *Language and Law/Linguagem e Direito* 9: 9–24. [CrossRef]

Prieto, Pilar, and Maria Teresa Espinal. 2020. Negation, prosody, and gesture. In *Oxford Handbook of Negation*. Edited by Viviane Déprez and Maria Teresa Espinal. Oxford: Oxford University Press, pp. 1–20.

Prieto, Pilar, Mariapaola D'Imperio, and Barbara Gili Fivela. 2005. Pitch accent alignment in Romance: Primary and secondary associations with metrical structure. *Language and Speech* 48: 359–96. [CrossRef]

Ripley, David. 2020. Denial. In *Oxford Handbook of Negation*. Edited by Viviane Déprez and Maria Teresa Espinal. Oxford: Oxford University Press, pp. 1–13.

Roitman, Malin. 2017. Introduction. In *The Pragmatics of Negation*. Edited by Malin Roitman. Amsterdam: John Benjamins, pp. 1–14.

Soldner, Felix, Verónica Pérez-Rosas, and Rada Mihalcea. 2019. Box of lies: Multimodal deception detection in dialogues. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Edited by Jill Burstein, Christy Doran and Thamar Solorio. Minneapolis, MN: Association for Computational Linguistics, pp. 1768–77.

United Nations. 2013. Statement Submitted by the Academic Council on the United Nations System, a Non-Governmental Organization in Consultative Status with the Economic and Social Council. Vienna Declaration on Femicide. Available online: https://www.unodc.org/documents/commissions/CCPCJ/CCPCJ_Sessions/CCPCJ_22/_E-CN15-2013-NGO1/E-CN15-2013-NGO1_E.pdf (accessed on 20 July 2024).

van der Sandt, Rob. 1991. Denial. *Chicago Linguistic Society* 27: 331–44.

Vandamme, Fernand. 1972. On Negation: An Interdisciplinary study. *Logique et Analyse* 15: 39–101.

Vrij, Aldert, Gün Refik Semin, and Ray Bull. 1996. Insight into behavior displayed during deception. *Human Communication Research* 22: 544–62. [CrossRef]

Wang, Min. 2024. Stance-taking in American courtroom interaction from the dual perspectives of language and gesture: A case study of the trial of 'The State of Minnesota v. Derek Michael Chauvin 2021'. *The International Journal of Speech, Language and the Law* 31: 162–72. [CrossRef]

languages

MDPI

Article

Forensic Audio and Voice Analysis: TV Series Reinforce False Popular Beliefs

Emmanuel Ferragne [1,*], Anne Guyot Talbot [1], Margaux Cecchini [1], Martine Beugnet [2], Emmanuelle Delanoë-Brun [2], Laurianne Georgeton [3], Christophe Stécoli [3], Jean-François Bonastre [4] and Corinne Fredouille [4]

[1] Laboratoire CLILLAC-ARP, UFR d'Études Anglophones, Faculté Sociétés et Humanités, Université Paris Cité, 75013 Paris, France; anne.talbot@u-paris.fr (A.G.T.); margaux.cecchini@etu.u-paris.fr (M.C.)
[2] Laboratoire LARCA, UFR d'Études Anglophones, Faculté Sociétés et Humanités, Université Paris Cité, 75013 Paris, France; martine.beugnet@u-paris.fr (M.B.); delanoee@u-paris.fr (E.D.-B.)
[3] Laboratoire Central de Criminalistique Numérique, Service National de Police Scientifique (SNPS), 69130 Écully, France
[4] Laboratoire Informatique d'Avignon, Avignon Université, 84000 Avignon, France; jean-francois.bonastre@univ-avignon.fr (J.-F.B.); corinne.fredouille@univ-avignon.fr (C.F.)
* Correspondence: emmanuel.ferragne@u-paris.fr

Abstract: People's perception of forensic evidence is greatly influenced by crime TV series. The analysis of the human voice is no exception. However, unlike fingerprints—with which fiction and popular beliefs draw an incorrect parallel—the human voice varies according to many factors, can be altered deliberately, and its potential uniqueness has yet to be proven. Starting with a cursory examination of landmarks in forensic voice analysis that exemplify how the voiceprint fallacy came about and why people think they can recognize people's voices, we then provide a thorough inspection of over 100 excerpts from TV series. Through this analysis, we seek to characterize the narrative and aesthetic processes that fashion our perception of scientific evidence when it comes to identifying somebody based on voice analysis. These processes converge to exaggerate the reliability of forensic voice analysis. We complement our examination with plausibility ratings of a subset of excerpts. We claim that these biased representations have led to a situation where, even today, one of the main challenges faced by forensic voice specialists is to convince trial jurors, judges, lawyers, and police officers that forensic voice comparison can by no means give the sort of straightforward answers that fingerprints or DNA permit.

Keywords: forensic phonetics; speaker identification; voice comparison; TV series

Citation: Ferragne, Emmanuel, Anne Guyot Talbot, Margaux Cecchini, Martine Beugnet, Emmanuelle Delanoë-Brun, Laurianne Georgeton, Christophe Stécoli, Jean-François Bonastre, and Corinne Fredouille. 2024. Forensic Audio and Voice Analysis: TV Series Reinforce False Popular Beliefs. *Languages* 9: 55. https://doi.org/10.3390/languages9020055

Academic Editors: Julien Longhi and Nadia Makouar

Received: 10 December 2023
Revised: 22 January 2024
Accepted: 24 January 2024
Published: 2 February 2024

1. Introduction

1.1. Background and Goals

Although firm quantitative evidence is still sparse (Eatley et al. 2018), it is often assumed that the many popular TV shows revolving around forensic science that emerged in the early 2000s (e.g., the various versions of the *Crime Scene Investigation* (*CSI*) franchise) have triggered what is now known as the "CSI Effect". To give but one example supporting this assumption, Call et al. (2013) surveyed 60 jurors from five malicious wounding juries in the United States. Their findings show that 95% of the jurors watched *CSI*, and 73% considered that the series influenced their verdict. The CSI effect is characterized by at least two phenomena (Eatley et al. 2018): (i) jurors in trials now tend to have unrealistic expectations regarding the presence of scientific evidence, and (ii) when scientific evidence is available, they are more prone to view it as infallible. This effect possibly extends beyond the general public and is thought to affect professionals (Call et al. 2013; Trainum 2019) and even criminals (Baranowski et al. 2018).

One of our central claims in the current article is that forensic voice analysis lends itself particularly well to misconceptions, which are, in turn, promoted by TV crime shows.

What is it that makes forensic voice analysis so different from other biometric evidence like fingerprints or DNA? Firstly, instances of forensic voice analyses have been comparatively sporadic in France: over the last 12 years, *Service National de Police Scientifique* (SNPS: national police forensic department) has performed 233 such analyses, while in August 2022 only, they carried out as many as 10,000 fingerprint and 7000 DNA analyses. Consequently, most courts have little to no experience in voice analysis, and fictional representations may sometimes be their only available reference. Secondly, while most people have never performed fingerprint or DNA analyses, they have constantly relied on their ears to identify people around them. Therefore, we all have strong intuitions based on experience, and we may have implicitly made the wrong generalization that auditory speaker identification is reliable under all circumstances. Thirdly, the analogy with fingerprints, which is constantly reinforced by works of fiction and the media, is deeply rooted in people's minds to the extent that they forget how variable and alterable voices can be.

The goal of this article is to analyze the representation of forensic voice analysis in a set of popular TV procedurals that have been broadcast in France. We focus on English-speaking, mainly US-based, shows since the latter are ubiquitous on French TV. We follow a mixed-methods approach combining qualitative analyses partially inspired by the field of visual studies with quantitative data to support our findings. A small-scale experiment involving plausibility ratings complements the analysis. The originality of our approach is strengthened by the diversity of the authors' backgrounds: there are three phoneticians and two specialists of speech processing who have been involved to varying degrees in forensic voice analysis over the last 5 (for the relative newcomers) to 30 years. Two authors specialize in visual studies, especially cinema and television studies, with one of them focusing on crime TV shows. The two forensic scientists from the audio section of SNPS are currently the only two scientists who perform forensic voice analysis in the police force in France, and they, therefore, provide first-hand experience and knowledge. In the fields of phonetics and voice processing, our areas of specialization cover a broad spectrum that encompasses most (if not all) fields related to forensic audio: perceptual, acoustic, and articulatory phonetics, automatic speech processing and recognition, and speaker identification. The next section briefly reviews historical landmarks that illustrate the two popular beliefs on which the article focuses: human listeners' ability to identify people by their voices and the idea that voices are as unique and unalterable as fingerprints (the voiceprint fallacy). Then, we briefly describe the specific context of this study: forensic voice analysis in France after a more thorough review of the situation at the international level. The rest of the article shows how we collected and annotated our data, and the type of qualitative and quantitative analyses that were performed.

1.2. Some Landmarks in Forensic Voice Analysis

About 90 years ago, Bruno Hauptmann was executed in the US for kidnapping and murdering Charles Lindbergh's two-year-old son (Solan and Tiersma 2003). Three years earlier, as Lindbergh and another man were delivering the ransom to the kidnappers in a cemetery, Lindbergh, who had remained in the car 60 to 100 m away, overheard the words "Hey, doctor. Here, doctor, over here" spoken with a German accent. Over two years later, Lindbergh confessed that it would be very difficult for him to recognize the man by his voice. Then, a district attorney asked Lindbergh if he wanted to see the man who killed his son; Hauptmann was brought in and asked to pronounce "Hey, doctor. Here, doctor, over here". Lindbergh said he recognized the voice he had heard in the cemetery. This case has remained controversial and has triggered a host of research on the many limitations of earwitness identification and voice parade methodology (see, e.g., Humble et al. 2022; McDougall et al. 2016; Yarmey et al. 1994), starting with McGehee (1937).

Thirty years after the controversial outcome of the Lindbergh case following questionable auditory voice identification methods, science came to the rescue: in an article entitled *Voiceprint identification*, Kersta (1962) suggested that one can identify people with a spectrographic analysis of their voice, and explained that very much like "people's finger-

prints, voiceprint identification uses the unique features found in their utterances". The original voiceprint method involved the visual analysis of the overall shape of spectrograms of ten frequent English words. Incidentally, Kersta's spectrograms came in two varieties: the broadband spectrogram, which is today's standard for phonetic description, and a contour plot that had better amplitude resolution than the other method at that time. Quite interestingly, contour plots are evocative of fingerprint ridge patterns, which may have strengthened the analogy between voice and fingerprints when Kersta's research was published.

In 2002, Elodie Kulik was raped and murdered after being involved in what seems to be a deliberate car accident caused by another car. Just after the crash, Elodie Kulik called the emergency services. On the bad-quality 26-s recording of this call, during which she realizes that the people who came to her were not here to help her, at least two male voices can be heard. Once the first man, who had died shortly after the events, was formally identified thanks to DNA evidence, a friend of his, Willy Bardon, was arrested as the potential second man. Bardon was eventually sentenced to 30 years of prison in 2021. One key aspect of this case is that some of Bardon's friends and relatives thought they had recognized his voice on the recording. His nephew says the voice on the recording displays "intonations" that sound like his uncle's. Even Bardon himself concurs: "It's my voice, it sounds like my voice; but I wasn't there." (Guiho 2020).

The Lindbergh and Kulik cases illustrate that while 90 years of research into our ability to auditorily identify someone by their voice have furthered our understanding of the limitations of such approaches, asking people if they can recognize someone's voice has remained a classic of police interviews and trial testimonies. The collective belief that people can definitely identify others by listening to their voice is firmly established, and for good reason: we rely on our ears to identify (or confirm the identity of) people around us on a daily basis.

Although Kersta's claims were soon disproven (Bolt et al. 1969), the voiceprint fallacy has lingered ever since then. For the anecdote, even the French researchers and engineers who worked with the Voice Identification Inc., *Sound Spectrograph*, *model 700* (before the generalization of personal computers for speech analysis) would borrow the English word and call it "le voiceprint". Note that it was only in 2007 that the International Association for Forensic Phonetics & Acoustics passed a resolution banning the use of the voiceprint approach.

1.3. Forensic Voice Analysis and Voice Comparison in the World Today

Forensic voice analysis involves a number of disciplines and tasks (signal processing, acoustic and perceptual phonetics, transcription of what is being said, etc.), and a comprehensive state-of-the-art section is well beyond the scope of the current article (see, e.g., De Jong-Lendle 2022; Hudson et al. 2021; Morrison and Thompson 2017; Watt and Brown 2020). However, some basic knowledge is needed so that readers can more fully appreciate the similarities and divergences between reality and fiction. One particular subfield that has generated much scientific debate and constitutes perhaps the most frequent type of analysis is voice comparison. The aim is to compare (at least) two recordings and determine how likely it is that they were spoken by the same person. There are four different methods, according to Morrison and Thompson (2017): auditory, spectrographic (this is equivalent to the voiceprint technique), acoustic–phonetic, and automatic. These methods are frequently combined, and some of these generic terms can be split into more fine-grained categories (e.g., acoustic analysis with or without statistical modeling).

The various studies surveying international practices over the past 20 years (e.g., Broeders 2001; Gold and French 2011, 2019; Morrison et al. 2016) show great variation between countries and practitioners. Some of these discrepancies stem from differences in national legal frameworks; others are the result of practitioners' habits and preferences. The evolution in recent years shows a general move towards standardization and methods that more closely match the requirements of general science. For example, the need for

reproducible results has led to the increasing use of automatic methods: between 2011 and 2019, the percentage of forensic experts in the world who use automatic speaker recognition systems went from 17% to over 40% (Gold and French 2019). Following the same paradigm shift towards reinforced scientificity, more and more practitioners use the Bayesian statistical framework and likelihood ratios in their reports rather than binary decisions or probabilities. Awareness of potential cognitive bias is yet another sign of the evolution of forensic science towards more controlled scientific methods (Cooper and Meterko 2019; Gold and French 2019).

However, discrepancies remain. One particular source of variability among the 39 laboratories surveyed in Gold and French (2011) is the number of cases: it ranged from 4 to 6000 with a mean of 506 (compare with France in Section 1.4). Another factor affecting variation between countries is the specific legal framework and, in particular, the admissibility of expert testimony. For instance, in the US, Federal Rule of Evidence 702 and the Daubert standard offer explicit criteria to determine admissibility (Morrison and Thompson 2017), while some other countries do not provide such explicit guidelines.

As far as methods are concerned, and in spite of the general trend towards standardized, thoroughly tested procedures, the latest surveys, by Morrison et al. (2016) and Gold and French (2019), illustrate that no unified methodology has emerged yet. For instance, in Morrison et al. (2016), while the auditory acoustic phonetic method was the preferred approach in Europe, all other possible approaches (including auditory-only and spectrographic) were used.

Other discrepancies between countries or forensic laboratories arise from differences in how they express their conclusions. The most frequent conclusion framework in Gold and French (2019) is the verbal likelihood ratio. It is surprising to note that around 5% of those surveyed still use binary conclusions (the criminal and the suspect are the same person or not).

In a word, it would be difficult to summarize what the state of the art in forensic voice analysis and voice comparison is because (i) practitioners' habits and legal frameworks are quite variable and (ii) the field encompasses many disciplines (e.g., phonetics, psychology, denoising, automatic speaker recognition, etc.) with their own performance tests, internal debates, etc. A recurring question that we are asked very often is how reliable automatic systems are. A reasonable answer is that it depends on the particular conditions. The following common sense example will illustrate this. In a white paper (Nuance Communications 2015), a company specialized in automatic voice authentication (e.g., a customer accesses online services provided by their bank by speaking a password that is then compared to a stored recording of this password by the customer) claims that the system can achieve 99% accuracy. The bank stores a finite set of pre-recorded utterances, the customers were cooperative when they recorded them, they probably make every effort to be as "recognizable" as possible each time they utter their password, and to be intelligible (e.g., they probably speak close to the telephone, at a volume that will not cause distortions, they avoid background noises, etc.). Now consider a realistic forensic scenario: the criminal's voice was captured by a microphone placed at a distance, drowned in background noise, and heavily distorted. Then, the forensic specialist organizes an interview with a suspect in order to record as much spoken material as possible so as to collect a reliable "known" sample for voice comparison. But the suspect will not cooperate and only provides one-word replies, or maybe they deliberately change their voice or their accent. If we then factor in that, contrary to the bank example, the number of "customers" is now potentially unlimited, it is easy to understand that accuracy levels drop dramatically. Morrison and Thompson (2017) argue convincingly that the admissibility of any approach in forensic voice analysis depends on "whether it has been empirically tested under conditions reflecting those of the particular case under investigation, and found to be sufficiently valid and reliable". In other words, reliability and performance should be assessed on a case-by-case basis.

1.4. Forensic Voice Comparison in France

What we describe in this article applies to a particular context: French police officers, forensic scientists, and viewers watching crime dramas shot (mainly) in the US. Background information is, therefore, necessary to clarify the context against which our analysis has been performed. At the time of writing this article, among the French police and gendarmerie forces, only the audio section of SNPS—basically the two co-authors whose affiliation is SNPS—perform forensic audio analysis and voice comparison in France. They use the human-supervised automatic approach complemented by the acoustic–phonetic method. No explicit criteria for the admissibility of evidence in court (like Rule 702 in the US) exist in the French system. The conclusions of an automatic voice comparison are expressed as a verbal likelihood ratio with an interpretative 11-point scale with five degrees reflecting the strength of the difference between the two samples, one neutral degree, and five steps indicating the degree of similarity between the two samples.

Through the Association Francophone de la Communication Parlée (AFCP—French Association for Spoken Communication) and its predecessors, French academics specializing in audio voice processing and phonetics passed a motion in 1990 (and again in 1997 and 2002) warning against the weaknesses involved in trying to identify someone by their voice given the current state of knowledge. In addition, the motion insists that academics should not perform forensic voice comparison, and overall, academics have complied with the injunction ever since it was accepted. This is in stark contrast to international practices: Gold and French (2011), in their survey, had 18 voice forensic practitioners affiliated with universities or research institutes out of their 36 respondents (the number dropped to 8 out of 39 in Gold and French 2019).

The history of forensic voice comparison in France over the past 30 years has been rather tumultuous (Boë 2000; Bonastre 2020). Relationships between academics, forensic scientists from law enforcement agencies, and private labs with self-styled audio experts have been, at times, very tense indeed. With the growing number of projects involving audio scientists from SNPS and academic phoneticians and specialists in signal processing over the last few years, collaborations have become fruitful, which is bound to have positive effects on the whole field.

As noted in the Introduction, not only have cases involving speaker identification been rare (compared to fingerprints or DNA) in France but also cases that have attracted national media attention are few and far between. In the last 15 years or so, the following cases can be mentioned: Benalla-Crase, Cahuzac, Chikli, etc. The iconic Chikli case has inspired movies and documentaries (e.g., Ratliff 2022) and may, therefore, have had a great impact on the public. Gilbert Chikli invented the CEO scam: he would call employees of large companies, pretend he was their CEO, and have money transferred for him. He even went so far as to pass himself off as the French Defense Minister for the same purposes.

We contend that the scarcity of cases of speaker identification that have hit the headlines in France, together with the low number of actual cases (relative to fingerprints or DNA), lead to a situation where judges, lawyers, police officers, and the public are not prepared to deal with such cases and may, therefore, rely on fictional representations.

2. Methods

2.1. Database Collection

The Springfield! Springfield! Website[1]—which was only accessible via the Internet Archive when we carried out this research—contains orthographic transcriptions of the dialogues of several thousands of English-speaking TV series and films. We manually explored the available titles, predominantly those of the crime genre, and selected 50 series that have been broadcast in France on non-encrypted, freely available TV channels. Our initial database comprises 5372 episodes, totaling 26,782,309 words.

We then wrote a MATLAB script to extract concordances targeting the word "voice", along with 80 characters of context before and after it. The 3321 occurrences of "voice" and their immediate lexical context were then individually inspected, and those that

seemed to be related to forensic audio analysis, and in particular to the identification of an individual by their voice, were kept. Some of the 160 passages we had thus identified were excluded after viewing them because they turned out to be irrelevant to our goals. Additionally, two shows, *Drop Dead Diva* and *Law and Order: UK*, were not accessible through the platforms we used: Prime Video and FMovies.

In total, 106 scenes from 28 different series were thus identified and extracted as video files. They were viewed and manually annotated based on several criteria:

- The episode, season number, and year of the first broadcast;
- Did the excerpt involve speaker identification?
- What methods were used?
- Was the analysis auditory and/or supported by visualizations of the signal?
- What type of visualizations were used;
- Was the display actually used or just decorative?
- Whether there was an explicit analogy with DNA;
- Whether there was an explicit analogy with fingerprints

Since this annotation stage was manual, we also collected more descriptive data: portions of dialogues that caught our attention, legitimate and erroneous uses of technical terms, and special descriptions of the graphical interfaces of signal processing programs.

The titles of the 28 TV shows with colors representing the number of episodes from each series in the dataset are shown in Figure 1.

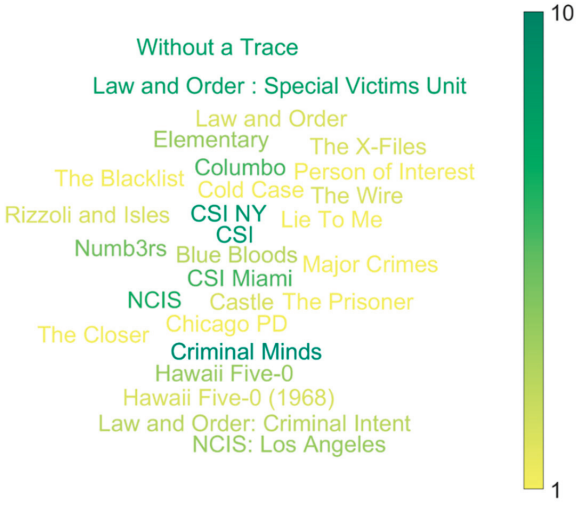


Figure 1. TV show title with colors indicating the number of excerpts.

Figure 2 shows the distribution of excerpts according to year. Most of them were produced during the *CSI* era; the last excerpt in the dataset first aired in 2016. The earliest scene was from an episode of *The Prisoner* from 1967: S01E10T21:57. In our notation, "S" introduces the season, "E", the episode, and "T" is followed by the time stamp corresponding to the beginning of the excerpt. Excerpt duration ranges from 15 s to 3 min and 15 s, with a mean of 59 and a standard deviation of 29 s for a total duration of 1 h and 45 min.

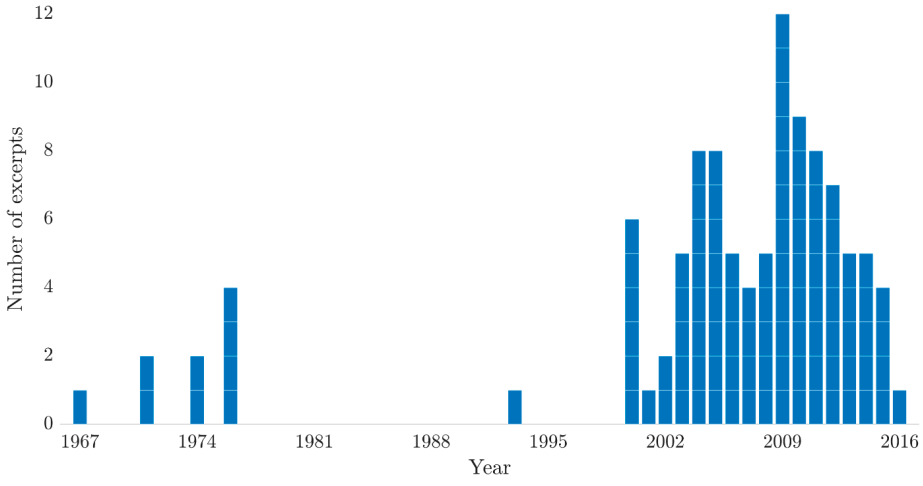


Figure 2. Distribution of excerpts according to year.

The claim that the TV shows we chose to study are omnipresent on French TV is substantiated by Figure 3. The graph shows the number of days when at least one episode of our series was broadcast on French TV between 1 January 2022 and 4 December 2023 (704 days). We wrote a MATLAB script to search a website that stores TV listings[2].

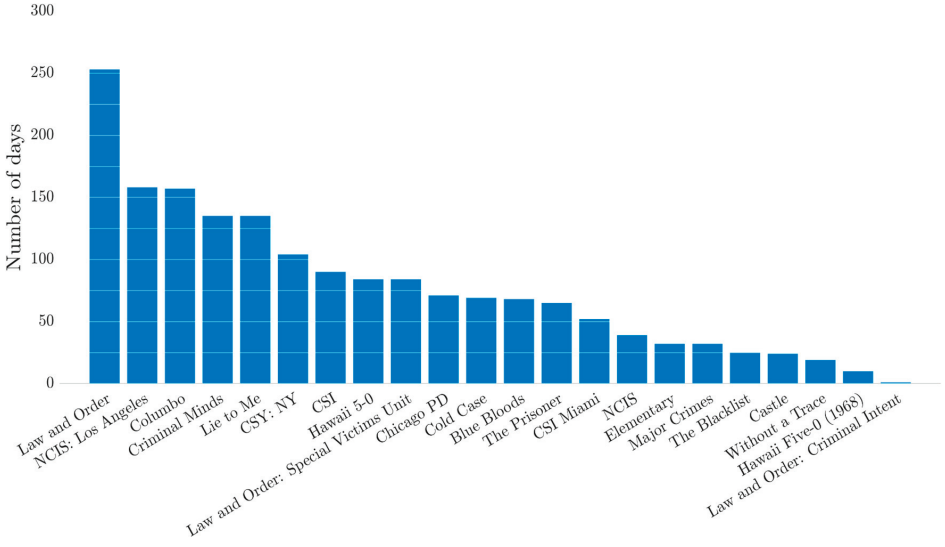


Figure 3. Number of days when at least one episode of the show on the x-axis was broadcast in France between 1 January 2022 and 4 December 2023.

2.2. Plausibility Ratings

The two forensic scientists from SNPS, as well as one speech-processing specialist from our team, rated the plausibility of 41 chosen excerpts that included visualizations of the speech signal. The five-point scale had the following values: 1 = totally unlikely, 2 = rather not likely, 3 = do not know, 4 = rather likely, and 5 = totally plausible. The participants carried out the rating experiment at their own pace, in uncontrolled environments, using a spreadsheet file. They were free to add any comment they deemed relevant.

3. Results

3.1. Database Analysis

Table 1 offers a synoptic view of the main findings resulting from the manual annotation and classification of the video excerpts in our database.

Table 1. Frequency of particular features of interest in our database.

Features of Interest	Number of Excerpts (Out of 106)
Auditory identification of a speaker from an audio recording	22
Comparison of two recordings in order to identify the speaker	42
Audio signal is graphically represented	65
Superimposed waveforms	5
Comparison of two waveforms side by side	4
Decorative waveforms	27
Voice is compared to DNA	0
Voice is compared to fingerprints	3
Analysis of the speaker's accent	7

Voice comparison (here, strictly speaking, the acoustic comparison of two recordings) occurs in 42 excerpts. For 20 of these, the analysis takes place within the excerpt, while the others refer to past or future analyses. Among the different types of methods, we found 22 instances of auditory analyses in which an individual is identified by their voice, with 15 of them occurring within the excerpt. The remaining cases are quite diverse; they involve simply listening to a recording, sometimes employing signal enhancement techniques, filtering, and source separation, particularly to analyze background noise.

An extreme case is depicted in Figure 4: at the top of panel A, the signal comes from one speaker's voice, at the bottom is the voice of another speaker, and between them, the two perfectly identical waveforms are being overlaid in the middle panel of the interface, resulting in a perfect overlap in panel B, with the description: "MATCH 99.675%". All the authors of the current paper agree that this is unrealistic because (i) no matter how hard one may try to produce two identical versions of an utterance, the minute variations in air pressure that are reflected in the waveform are well beyond his or her control, leading to different waveforms, (ii) software would normally output likelihood ratios rather than raw percentages, and (iii) the human eye cannot detect speaker-specific information in such raw representations of the signal as waveforms. In our dataset, the comparison of two identical waveforms occurs on nine occasions: five of them have waveforms displayed side by side, and the remaining four are superimposed. We contend that this deceptive trick is particularly appealing because of its simplicity and its similarity with fingerprint analysis.

We found references to accents or dialects in seven episodes. For example, a voice sample is submitted to "a beta version of Shibboleth, [. . .] an accent identifier", a very aptly denominated fictional software program, in *NCIS: Los Angeles* S02E12T39:50. In *Law and Order: New York* S03E17T29:07 a 911 caller articulates his /t/ in the words *battery* and *city* as plosives rather than the flap consonants that are much more usual in American English. The voice expert in this excerpt regards this as an idiosyncrasy that, taken in conjunction with other similarities, does not constitute a positive ID but a probable match. In *Hawaii Five-0 (1968)* S08E23T18:26, the voice specialist says she "did pick up on a couple of flat A's and a dropped G in the word *tellin*", which leads her to conclude that the caller comes from the South West of the United States. While these examples are not too far-fetched—automatic accent identification in English is very accurate (Zuluaga-Gomez et al. 2023), and auditory accent identification has sometimes been used (e.g., S. Ellis 1994), their forensic value when it comes to identifying a single speaker is limited.

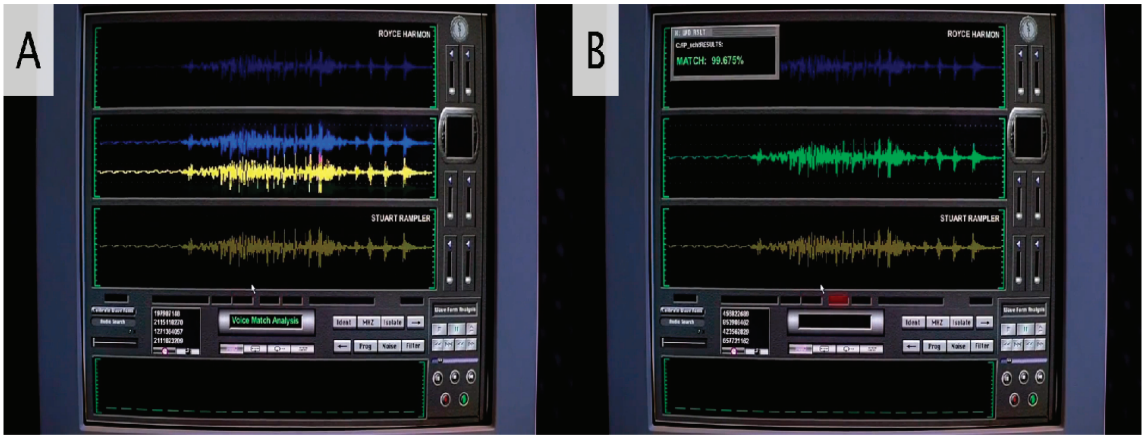


Figure 4. The (unrealistic) classic perfect matching of two waveforms (*CSI*: S01E08T18:39). The identical blue and yellow curves in (**A**) are gradually merging into a green one in (**B**).

Regarding the potential analogy of the voice with biometric data, a possible similarity with DNA is never mentioned in our dataset. An explicit comparison with fingerprints is present in three excerpts. In *The Prisoner* S01E10T21:57, Number Two explains that "Voices are like fingerprints; no two are the same. Even if the voice is disguised, the pattern doesn't change". In *CSI NY* S07E15T24:50, Detective Mac Taylor claims that "voice patterns are as distinct as fingerprints". In *Law and Order: New York* S14E16T29:59, a character says: "I heard your voice on the tape. It's like a fingerprint". These episodes first aired in 1967, 2011, and 2013, respectively, which shows that the use of the term extends well beyond the rejection of the concept by the academic community.

Additionally, there are three excerpts where the term "voiceprint" (with various spellings) is displayed on software graphical user interfaces; two of them are shown in Figure 5. Incidentally, the two screenshots bear a close resemblance to one another because they come from episodes of the same show that came out one year apart. In addition, the superposition of key acoustic cues is arguably reminiscent of a DNA electropherogram. Therefore, although DNA is never mentioned explicitly as an analogy, it is evoked through visual displays.

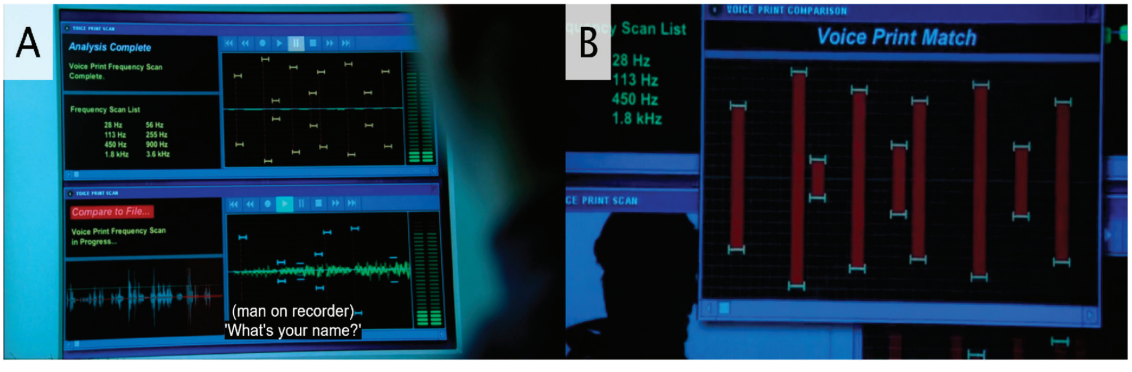


Figure 5. A visual evocative of a DNA electropherogram where "voice print" also appears. ((**A**) *Without a trace* S02E22T05:39; (**B**) *Without a trace* S01E18T29:45).

"Voiceprint" is explicitly mentioned in the dialogues of five excerpts. In *Criminal Minds* S11E09T28:41, Special Agent Jennifer Jareau threatens a woman that voiceprint recognition

will be applied to the recording of a phone call, and the woman immediately (rightly) opposes: "that's not foolproof". This is the only excerpt where a character's awareness of the limitation of the method is clearly stated.

Familiarity with a speaker's voice is sometimes used as a reason supporting speaker identification: in *NCIS* S01E09T16:14, a man claims he recognized, without a doubt, someone's voice on the answering machine because he and the caller have known each other since they "were second lieutenants at the Basic School". Similarly, in *Castle* S02E23T16:26, "even though they wore ski masks, he recognized them by their voices because they grew up together". The familiarity argument is taken a step further in *Criminal Minds* S01E05T15:22 where, as a kidnapper on the phone wants to speak to the twin sister of the woman he has kidnapped, Special Agent Elle Greenaway stands in, but the kidnapper says: "I know her voice therefore I know her sister's. Get off the phone". It is true that the voices of monozygotic twins tend to be more similar than those of genetically unrelated human beings (San Segundo and Künzel 2015). In *Criminal Minds* S03E12T03:29, familiarity with the voice is expected to be a robust predictor of successful identification ("the parents can ID the voice") and again in S04E14T19:24 (from the same show): a reproachful mother complains: "you think I don't know my own daughter's voice?". Our ability to recognize familiar voices has been shaped by evolution, it has reached a high degree of sophistication, and we rely on it to structure the world that surrounds us (Sidtis and Kreiman 2012). There is also compelling evidence that the processing of familiar and unfamiliar voices is distinct (Stevenage 2018). Therefore, in the excerpts we mentioned, scientific evidence supports, to a certain extent, the characters' claim that they can recognize familiar voices. But, as always with forensic voice analysis, this ability comes with a certain error rate that becomes worse as the sample length becomes shorter, its audio quality is more degraded (noisy or over the phone), and it may not be robust to voice disguise.

Visual representations of the audio signal appear in 65 of our excerpts. Fifty-two of these are (or have as their dominant plots) amplitude–time graphs, i.e., waveforms. Five other cases show spectra or spectrograms, and for the remaining eight cases, the display features a combination of graphs and sometimes graphs whose nature is difficult to determine. It appears then that by far the most frequent plot is the waveform, which is, arguably, the least informative type of signal visualization when one is interested in voice and phonetic analysis. A small sample of visual representations of the signal, reflecting, among other things, technological developments in the history of speech analysis on TV, are presented in Figure 6. Panel A shows an oscilloscope that is contemporary with the excerpt (1976); it is used here as a simple visual cue signaling that some audio signal is being played back (very much as modern phone apps would). Panel B shows a software program from 1993, supposedly from the "voice biometrics lab at Georgetown", with an "oscilloscope" (the signal dimension that it displays remains unclear) and a spectrogram. Panel C—which is, incidentally, the third graphical interface with "voiceprint" on it—shows a curious type of speech waveform that is reminiscent of the kind of low-bit-depth signal of electrocardiograms. Panel D displays the state-of-the-art televisual voice analysis gear with touchscreen capabilities and translucent colorful waves that seem to populate the whole room.

The visualizations play various roles. In Figure 7, panel A, the flashy displays in the background are totally independent of the ongoing analysis by the two protagonists. Colorful curves seem to be serving a purely decorative purpose in 27 excerpts. In panel B, what the analyst is looking at on the screen in front of him is duplicated on the huge screens behind him. The waveforms are obviously only intended for the viewers, probably both as a mise en scène ploy to make the passage livelier and as a way to call viewers as witnesses.

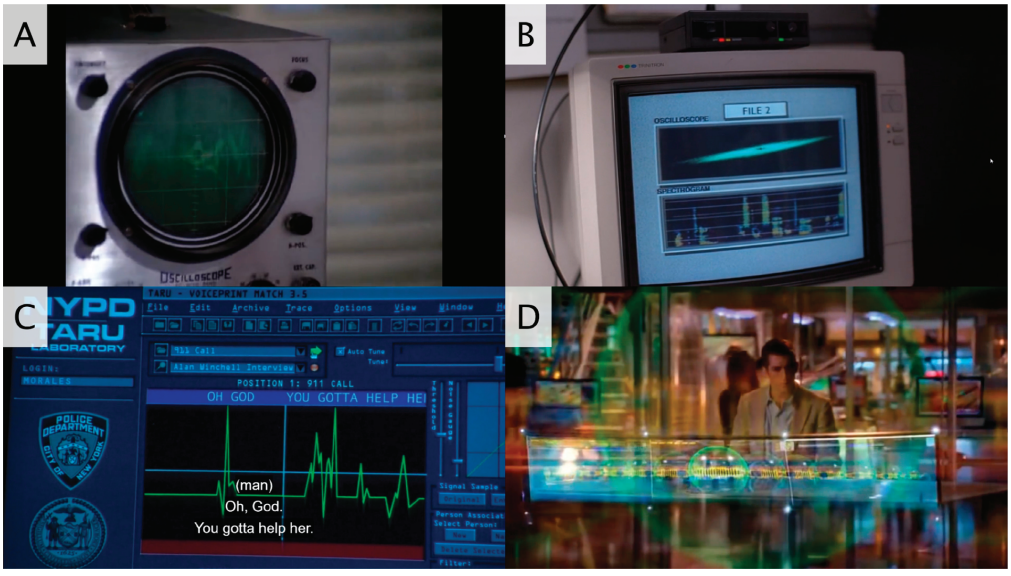


Figure 6. Various types of signal visualizations. (**A**) An oscillogram from *Hawaii Five-0* (1968) S08E23T18:26. (**B**) Oscilloscope and spectrogram from *The X-Files* S01E07T17:38. (**C**) A waveform from *Law and Order: Special Victims Unit* S07E15T03:34. (**D**) Waveforms from *CSI: Miami* S10E08T28:51.

Figure 7. (**A**) *CSI* S15E18T27:40. (**B**) *CSI NY* S05E22T07:15.

3.2. Plausibility Ratings

The distributions of plausibility ratings by rater are shown in Figure 8. The two SNPS forensic scientists in the author list appear as SNPS-1 and SNPS-2, and ACADEMIC is one of the speech-processing specialists from the authors. All distributions are skewed in that they tend to exhibit more low than high scores. A Kruskal–Wallis test shows a difference in median scores among the three raters ($\chi^2 = 6.62$, $p = 0.04$), which, after post hoc comparison, is due to lower scores by SNPS-2 compared to ACADEMIC. Cohen's κ analysis shows that only the judgments of the two SNPS members exhibit significant consistency (SNPS-1~SNPS-2: $\kappa = 0.307$, $p < 0.01$; SNPS-1~ACADEMIC: $\kappa = 0.126$, $p > 0.05$; SNPS-2~ACADEMIC: $\kappa = 0.078$, $p > 0.05$). By-rater mean ratings are ACADEMIC: 2.68, SNPS-2: 1.98, and SNPS-1: 2.44. By-excerpt median scores show that only 8 of them score 4 or 5 (rather likely/totally plausible), while 28 score 2 or 1 (rather not likely/totally unlikely).

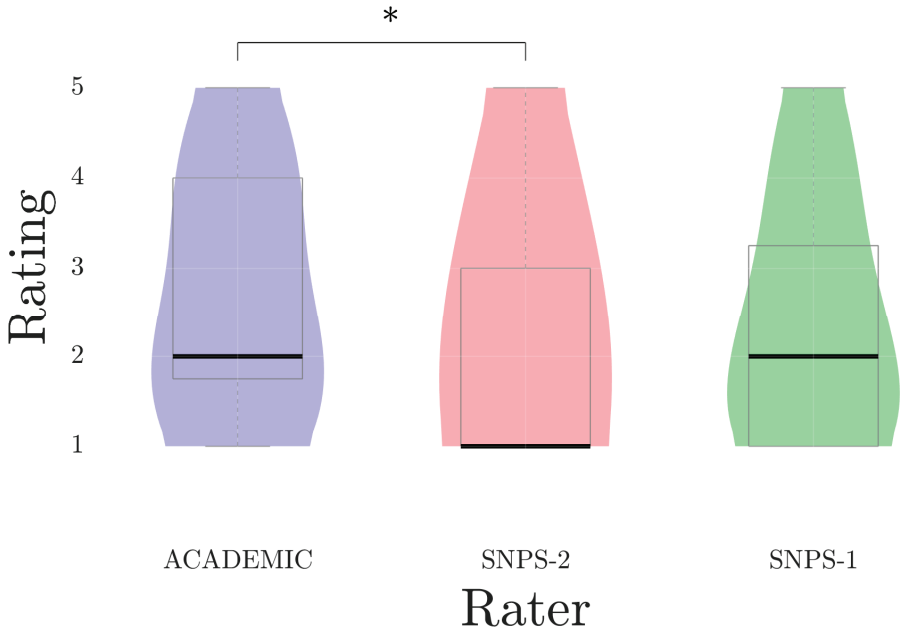


Figure 8. Plausibility ratings for each rater; *: statistically significant at the $p < 0.05$ level.

A quick look at the raters' comments shows that the disagreement between fiction and reality may stem from several reasons. SNPS-1 asserts that the visual displays are hardly ever credible. This discrepancy is particularly marked when voice comparison is concerned, with TV series showing, e.g., flashing and superimposed waveforms while the software used at SNPS outputs tables with likelihood ratios and unspectacular curves. Some differences may be the result of country-dependent rules: SNPS-1 and SNPS-2 remark that the French judicial system would not allow the recording of a person in a police interview without the person being informed prior to the interview. For example, in *Without a Trace* S02E22T31:30, as soon as the character enters the room, she declares, "I'm not sure what I can tell you, but if you think I can help…". Her words are instantly recorded, and within the next 10 s, her voice is compared with that of an unknown caller, and the message "Voice Print Match 100%" flashes on the computer screen. The overall feeling shared by all three raters is that the signal processing techniques used in the excerpts are usually too good to be true. Audio enhancement and source separation give much better results than what one would expect in real-life situations. The time dimension is nearly always unrealistic. While real-life forensic scientists spend long hours listening to and transcribing audio recordings, their fictional counterparts complete the job in a split second. This is the case in the aforementioned scene from *Without a Trace*. The scene in *CSI* S06E12T26:05 follows the same pattern when Forensic Scientist Nick Stokes asks another scientist: "if you've got a couple of minutes, I need a voice comparison". Another recurring critique concerns the exaggerated use of signal visualizations, as demonstrated in the previous section.

3.3. More on Aesthetics

In the context of TV shows, where visibility is key, voice analysis, as performed by experts, pertains to the immaterial. For a viewer accustomed to seeing clues, bodies, or bullets, visual representations of audio signals resolve the challenge posed by sound, invisible by nature, in a medium where the image takes precedence (Chion 2003). Non-figurative visualizations belong to a category of images commonly referred to as operational or operative (Hoel 2018). Even when they stem from the "remediation" (Bolter and Grusin 1999) of an older medium (such as analog audio recordings), operative images that are associated with

automated systems are no longer designed for the human eye; they participate in a mode of representation that depends neither on the human scale nor perspective (Hoel 2018).

In police procedural drama shows, the presence of these images emerges from a dual discursive regime that merges belief and scientific expertise, manifesting itself in staging, scientific discourse, and character typification (with a dark room and geeky technician). The operative image, sometimes the main source of light, is the basis for an explanatory exchange, where the description of processes and scientific data allows the expert to demonstrate mastery of specialized terminology.

3.4. Specialized Terminology

One aspect that makes the selected scenes potentially more convincing is the use of technical terms that are, at least partially, correct. In *The X-Files* S01E07T17:38, Special Agent Fox Mulder, after being shown two spectrograms exhibiting a near-perfect match, states, "He may have disguised his voice electronically, but he couldn't alter the formants unique to his own speech patterns". Of course, "formants" is a technical term that refers to frequency bands with high energy in the spectrogram. Now, as to whether one can alter one's formant pattern: clearly, at least the lowest three formants, those that are used to form speech sounds, can easily be altered without the help of technology. And it is reasonable to say that all formants can be altered electronically.

In *NCIS* S01E09T16:14, Forensic Specialist Abby Sciuto explains that "Ma Bell eliminates any frequency that's below 400 Hz and above 3400: it allows for longer distance transmission". To the best of our knowledge, this is accurate, but Abby's credibility quickly evaporates when, seconds later, her computer screen displays perfectly matching waveforms that lead her to conclude that they both come from the same speaker.

In *Law and Order* S11E08T07:30, the forensic specialist claims that "the average grown male has a pitch frequency of 130 Hz; a teenage boy post-puberty is about 140; this one is at 152". When detective Lennie Briscoe objects: "How does that make him a teenager?" the audio specialist replies: "People go up 10 to 15 Hertz when they're screaming". The reference values here do not seem far-fetched, and the audio specialist cautiously mentions at one point that this is just an educated guess. However, these are clearly mean values, and these reference pitch values would be of very limited use in authentic forensic contexts, given the range of within and between-speaker variation. For example, in a study involving 100 male speakers of British English aged 18–25 years old, mean individual pitch in spontaneous conversations ranges from about 85 Hz to about 140 Hz (Hudson et al. 2007).

From these three (and other) examples in our dataset, it appears that the technical jargon is not necessarily used to deceive viewers but is not to be blindly trusted: both accurate and inaccurate technical vocabulary and facts can occur within very short time windows. As experts, the authors can scrutinize techniques and jargon with a critical eye, but for the lay observer, discriminating between veracity and implausibility is challenging. The fictional aspects, exacerbated by an improbable timeline and the ease with which the unfolding of events occurs, nevertheless yield a cohesive construct to most viewers.

3.5. Lab Technicians

While technical terminology is linked to the sanctuary (a dark room illuminated by artificial light sources), the signal analysis expert is often no ordinary individual. A lonely IT genius navigating a parallel digital world of lavish sartorial opulence or Gothic fashion style, a gifted musician who "hears in perfect pitch" (*CSI* S01E08T17:20), characters with technical expertise handle seemingly incomprehensible waveforms, akin to gurus conveying an enigmatic message. Figure 9 shows a sample of forensic analysts who have eccentric behaviors or outfits. In panel A, Sherlock Holmes, who performs voice analysis himself here, is a recovering drug addict. In panel B, Forensic Scientist Abby Sciuto mixes gothic fashion with a formal lab coat. In C, the forensic specialist is a cliché Black American gifted musician with the stereotypical dress code, language, and tendency to flirt; his name says it all: Disco Placid. And in D, Penelope Garcia is yet another nonconformist analyst

who, besides being famous for her colorful clothes and eccentric behavior, is a former hacker who was caught by the FBI and given the choice to either spend the rest of her life in prison or work for them.

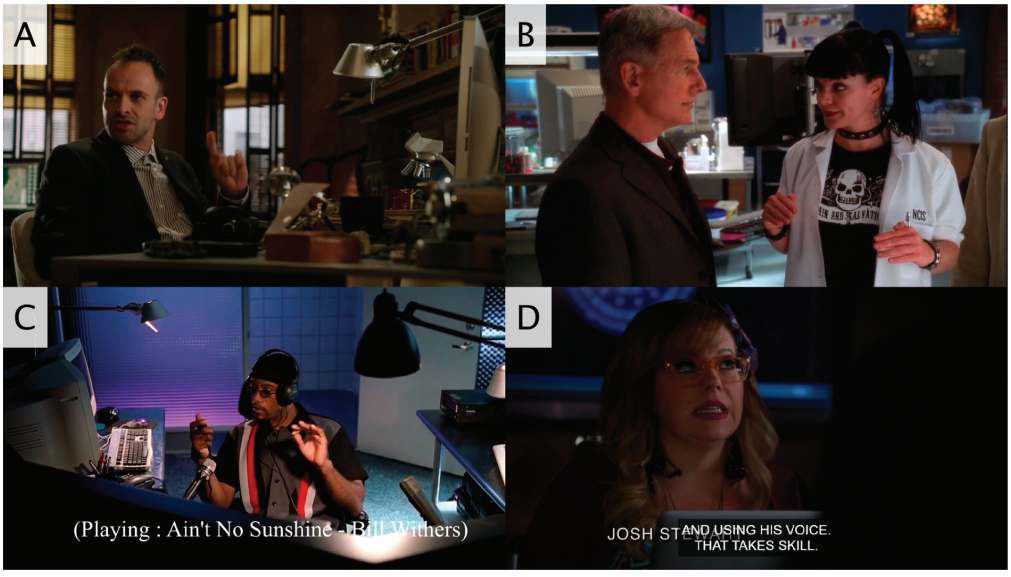


Figure 9. (**A**) *Elementary* S02E15T28:57. (**B**) *NCIS* S04E19T17:28. (**C**) *CSI* S01E08T17:20. (**D**) *Criminal Minds* S11E22T6:32.

4. Discussion

Forensic voice analysis in TV series shows great variation both in terms of techniques (acoustic, auditory, automatic or not, etc.) and degree of plausibility. From the grotesque perfect match between two waveforms to more moderate (and realistic) opinions and statements, popular crime procedurals mix science and entertainment. According to Kirby (2017), we live in the golden age of the fusion between the two, and our analysis confirms this: scientific terminology is applied to realistic and unrealistic contexts alike; scientific-looking waves are constrained to perform unscientific tricks, etc. Even in a recent serious podcast entirely devoted to one of us describing his profession as a forensic voice specialist, dramatic music effects and descriptions that are typical of detective books were used to glamourize the story.

Speech visualization in detective series is a recurrent motif. In addition to the static presence of more conventional items (recorders, cassette tapes, USB keys, etc.) and traditional narrative strategies (close-up shots on a silent character whose face reflects an effort of attention), speech signal visualizations substitute a dynamic solution which gives voice analysis a live and dramatically tangible dimension compared to simple listening. We noted that the preferred visual representation of the speech signal is the waveform, i.e., the amplitude-time graph, whereas it is our experience that for the analysis of speech, other visuals, like spectrograms, are much more informative. Intuitively, we feel that waveforms not only convey a "live" dimension thanks to their rapidly changing patterns, but they also favor the analogy with fingerprints. And the latter is made possible because it is easy to overlay two waveforms and let the viewer confirm that a perfect match has been found. The main difference, which is critical here, is that viewers know what fingerprints represent; they are figurative, i.e., they are a faithful representation of the thing they represent. Signal visualizations, on the contrary, are the result of a conventional synesthetic transformation that non-experts may not fully understand. It is, therefore, easy to trick people into believing that waveform matches are as robust as fingerprint matches.

The plausibility ratings we obtained from two forensic scientists and one academic specializing in speech processing show low values, supporting the overall lack of realism. However, the results are just preliminary. A tentative explanation for the lack of agreement between the academic and the two raters from SNPS is that the academic probably assessed plausibility with respect to what speech processing techniques would allow, whereas the SNPS colleagues evaluated plausibility against their actual professional practice. But only a more comprehensive rating scheme, with more raters and a more detailed set of instructions, would allow robust generalizations. In particular, the current ratings assess the plausibility of these excerpts against forensic audio analysis in France and the French judiciary. A panel of international forensic voice specialists would, therefore, be a useful addition to the current study.

The reception of American fictional crime shows by French viewers warrants a few comments. English-speaking TV series are systematically dubbed; it is impossible to quantify how many viewers switch to the original soundtrack, and we, therefore, cannot guarantee that they heard the exact terms we comment on in this article. A comparative study of the French and English versions of the dialogues constitutes an interesting potential follow-up. A remarkable by-product of dubbing is that it confirms the strong influence of US TV series on French audiences. French judges have been annoyed at being often called "Votre Honneur" (a calque from the English "Your Honor" that is very frequent in dubbed versions), and policemen are reportedly irritated when someone they have just arrested wants to make the phone call people in custody in American TV series generally make (Villez 2005).

The discrepancy between the primary intended target—the North American audience—and the secondary, French, viewership may have unexpected effects. *Law and Order*, for example, has a very local flavor: references to real events and criminal cases that took place in NYC are numerous, many actors from other shows have participated, two mayors (Bloomberg and Giuliani) have appeared as themselves, etc. In short, *Law and Order* has become an institution that reflects the local context (Villez 2014). French viewers are bound to miss a number of these allusions and references, resulting in an increased distance between the show and its audience once it has crossed the Atlantic. Le Saulnier (2012) found that the French police officers he surveyed preferred TV crime series whose setting was remote from their own professional setting. Such series allow them to drop their expert judgments and enjoy an action they now regard as plausible since, e.g., they do not specialize in the North American legal system.

As far as the CSI effect is concerned, our study does not go so far as to investigate a potential link between people's viewing habits and their faith in forensic evidence. Our aim was to offer an overview of the various on-screen representations of forensic voice analysis in order to examine what French jurors, police officers, and judges potentially have in mind when evidence based on voice analysis is presented to them. This by no means implies that the overstated efficiency of the techniques shown in TV series actually affects people. In fact, Ribeiro et al. (2019) studied the link between their participants' exposure to forensic science on television and their beliefs about the accuracy of various types of analyses. They did not find evidence that, as the CSI effect predicts, the more you are exposed, the more you trust these techniques. When comparing voice analysis to other techniques, such as DNA, toxicology, or blood pattern analysis (etc.), their participants responded that voice analysis was among the least accurate techniques and those that involve a high proportion of human judgment. Why this is the case is unclear to us at the moment, but perhaps we can assume that various efforts to vulgarize forensic science and voice analysis (Gully et al. 2022; Mauriello 2020; Smith 2023) have come to fruition, and we hope the current article will serve the same function and add to the existing body of knowledge.

Possible extensions include the collection of more recent TV shows since it appears that we are now in the post-CSI televisual age where the deductive (and fallible) reasoning that was typical of the pre-CSI series has been resurrected (Bull 2016). And quite logically, a study of forensic voice analysis in French crime series would be very informative. The

extension to the big screen would also be welcome since the various screen sizes, from the cinema to mobile phones, do not imply the same constraints to captivate the viewers (Beugnet 2022; J. Ellis 2006). Such studies would be all the more useful as Rafter (2007) has noted that contemporary fiction contributes to developing, alongside professional criminology, a "popular criminology".

Many new challenges in forensic speaker identification and voice comparison have emerged. Some of them are the result of recent technological advancements. For example, voice cloning technology, which "impersonates" someone after learning this person's main vocal features from audio samples, yields outputs that are becoming more and more convincing. And beyond sheer fraud detection, these new technological possibilities pose social, ethical, and legal problems linked, among others, to intellectual property (Watt et al. 2020). Other challenges include the impact of the media, social media, and fiction and how we, as scientists and forensic specialists, should disseminate our knowledge on voice analysis and speaker identification. The audio experts from SNPS, among the authors, insist that a sizeable part of their time is devoted to explaining the limitations of their work and that, contrary to what most people think, speaker identification and voice comparison should not be taken for granted. Now that, in recent years in France, audio specialists from SNPS have started collaborating with the academic world, one of the challenges for the near future is to maintain our efforts in this direction. In parallel, training new specialists and ensuring that forensic science complies with the basic rules of general science (transparency, peer review, replicability, etc.) are among our priorities.

5. Conclusions

The inspection of over 100 excerpts from (mostly) American TV series that portray forensic voice analysis showed that, more often than not, fiction exaggerates the possibilities of speech processing and the human ear. We expect that such fictional depictions favor the persistence of the voiceprint fallacy and the false belief that humans can identify people's voices reliably. The constraints inherent in entertainment make the various forensic techniques in TV crime fiction more efficient, more visual, and less time-consuming than their real-life counterparts. Plausibility ratings of our excerpts by two forensic audio specialists and one speech-processing researcher were very low overall. Given the relative scarcity of criminal cases involving speaker identification and voice comparison (at least in France), our default expectation is that the only representation that people (jurors, lawyers, judges, police officers, forensic scientists in other fields) have in mind come from fictional works, and TV series in particular. Here is a tentative description of the average televisual false representation of forensic voice analysis: lay people can infallibly recognize somebody's voice, especially if they are familiar with the speaker. No matter how intelligible the original audio signal is, the speaker's voice can easily be isolated from surrounding noises and enhanced if necessary. Televisual forensic experts, who tend to specialize in an unrealistically wide range of scientific disciplines and whose behavior and outfits are stereotypically eccentric, have suspects record exactly the same words as those on the questioned sample. Then, typically, strictly identical waveforms appear on a computer screen with a very high percentage supporting a "match" between the two voices. Now, back to reality, we will maintain our efforts to disseminate the type of educational content we have analyzed here, and we hope that other forensic voice specialists will use these examples to explain to others in what ways fictional depictions of forensic voice analysis may have biased their expectations. Beyond the study of fictional representations, we are quite confident that the general paradigm shift towards more scientifically validated methods in our field will increase the reliability of forensic voice analysis.

Author Contributions: Conceptualization, E.F. and A.G.T.; methodology, E.F. and A.G.T.; software, E.F.; validation, E.F., A.G.T., M.C., M.B., E.D.-B., L.G., C.S., J.-F.B. and C.F.; formal analysis, E.F., A.G.T., M.C., M.B., E.D.-B., L.G., C.S., J.-F.B. and C.F.; investigation, E.F., A.G.T., M.C., M.B., E.D.-B., L.G., C.S., J.-F.B. and C.F.; resources, E.F., A.G.T., M.C., M.B., E.D.-B., L.G., C.S., J.-F.B. and C.F.; data curation, M.C.; writing—original draft preparation, E.F. and A.G.T.; writing—review and editing, E.F.

and A.G.T.; visualization, E.F.; supervision, E.F.; project administration, E.F.; funding acquisition, E.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Idex Université de Paris, ANR-18-IDEX-0001: VoCSI-Telly-Émergence en Recherche.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data is copyrighted and cannot be shared.

Conflicts of Interest: The authors declare no conflict of interest.

Notes

1 https://www.springfieldspringfield.co.uk/ (accessed on 23 January 2024).

2 https://programme-tv.nouvelobs.com/programme-tv/ (accessed on 23 January 2024).

References

Baranowski, Andreas M., Anne Burkhardt, Elisabeth Czernik, and Heiko Hecht. 2018. The CSI-education effect: Do potential criminals benefit from forensic TV series? *International Journal of Law, Crime and Justice* 52: 86–97. [CrossRef]

Beugnet, Martine. 2022. The Gulliver effect: Screen size, scale and frame, from cinema to mobile phones. *New Review of Film and Television Studies* 20: 303–28. [CrossRef]

Boë, Louis-Jean. 2000. Forensic voice identification in France. *Speech Communication* 31: 205–24. [CrossRef]

Bolt, Richard H., Franklin S. Cooper, Edward E. David, Peter B. Denes, James M. Pickett, and Kenneth N. Stevens. 1969. Identification of a Speaker by Speech Spectrograms: How do scientists view its reliability for use as legal evidence? *Science* 166: 338–43. [CrossRef] [PubMed]

Bolter, Jay David, and Richard A. Grusin. 1999. *Remediation: Understanding New Media*. Cambridge, MA: MIT Press.

Bonastre, Jean-François. 2020. 1990–2020: Retours sur 30 ans d'échanges autour de l'identification de voix en milieu judiciaire. In *2e atelier Éthique et TRaitemeNt Automatique des Langues (ETeRNAL)*. Edited by Gilles Adda, Maxime Amblard and Karën Fort. pp. 38–47. Available online: https://aclanthology.org/2020.jeptalnrecital-eternal.5.pdf (accessed on 23 January 2024).

Broeders, Ton. 2001. Forensic Speech and Audio Analysis Forensic Linguistics 1998 to 2001. Paper presented at the 13th INTERPOL Forensic Science Symposium, Lyon, France, October 16–19; pp. 54–84.

Bull, Sofia. 2016. From crime lab to mind palace: Post-CSI forensics in *Sherlock*. *New Review of Film and Television Studies* 14: 324–44. [CrossRef]

Call, Corey, Amy K. Cook, John D. Reitzel, and Robyn D. McDougle. 2013. Seeing is believing: The CSI effect among jurors in malicious wounding cases. *Journal of Social, Behavioral, and Health Sciences* 7: 52–66.

Chion, Michel. 2003. *Un art sonore, le cinéma: Histoire, esthétique, poétique*. Paris: Cahiers du cinéma.

Cooper, Glinda S., and Vanessa Meterko. 2019. Cognitive bias research in forensic science: A systematic review. *Forensic Science International* 297: 35–46. [CrossRef]

De Jong-Lendle, Gea. 2022. Speaker Identification. In *Language as Evidence*. Edited by Victoria Guillén-Nieto and Dieter Stein. New York: Springer International Publishing, pp. 257–319. [CrossRef]

Eatley, Gordon, Harry H. Hueston, and Keith Price. 2018. A Meta-Analysis of the CSI Effect: The Impact of Popular Media on Jurors' Perception of Forensic Evidence. *Politics, Bureaucracy, and Justice* 5: 1–10.

Ellis, John. 2006. *Visible Fictions: Cinema, Television, Video (Nachdr.)*. London: Routledge.

Ellis, Stanley. 1994. The Yorkshire Ripper enquiry: Part I. *Forensic Linguistics* 1: 197–206. [CrossRef]

Gold, Erica, and Peter French. 2011. International Practices in Forensic Speaker Comparison. *International Journal of Speech, Language and the Law* 18: 293–307. [CrossRef]

Gold, Erica, and Peter French. 2019. International practices in forensic speaker comparisons: Second survey. *International Journal of Speech Language and the Law* 26: 1–20. [CrossRef]

Guiho, Mickaël. 2020. Willy Bardon condamné dans l'affaire Kulik: Les jurés expliquent leur décision. France 3 Hauts de France. Available online: https://france3-regions.francetvinfo.fr/hauts-de-france/somme/amiens/willy-bardon-condamne-affaire-kulik-jures-expliquent-leur-decision-1760827.html (accessed on 23 January 2024).

Gully, Amelia, Philip Harrison, Vincent Hughes, Richard Rhodes, and Jessica Wormald. 2022. How Voice Analysis Can Help Solve Crimes. *Frontiers for Young Minds* 10: 702664. [CrossRef]

Hoel, Aud Sissel. 2018. Operative Images. Inroads to a New Paradigm of Media Theory. In *Image—Action—Space*. Edited by Luisa Feiersinger, Kathrin Friedrich and Moritz Queisner. Berlin: De Gruyter, pp. 11–28. [CrossRef]

Hudson, Toby, Gea de Jong, Kirsty McDougall, Philip Harrison, and Francis Nolan. 2007. F0 Statistics for 100 Young Male Speakers of Standard Southern British English. Paper presented at the 16th International Congress of Phonetic Sciences: ICPhS XVI, Saarbrücken, Germany, August 6–10; pp. 1809–12. Available online: https://api.semanticscholar.org/CorpusID:17550455 (accessed on 23 January 2024).

Hudson, Toby, Kirsty McDougall, and Vincent Hughes. 2021. Forensic Phonetics. In *The Cambridge Handbook of Phonetics*, 1st ed. Edited by Rachael-Anne Knight and Jane Setter. Cambridge: Cambridge University Press, pp. 631–56. [CrossRef]
Humble, Denise, Stefan R. Schweinberger, Axel Mayer, Tim L. Jesgarzewsky, Christian Dobel, and Romi Zäske. 2022. The Jena Voice Learning and Memory Test (JVLMT): A standardized tool for assessing the ability to learn and recognize voices. *Behavior Research Methods* 55: 1352–71. [CrossRef]
Kersta, Lawrence G. 1962. Voiceprint Identification. *Nature* 196: 1253–57. [CrossRef]
Kirby, David A. 2017. *The Changing Popular Images of Science*. Edited by Kathleen H. Jamieson, Dan M. Kahan and Dietram A. Scheufele. Oxford: Oxford University Press, vol. 1. [CrossRef]
Le Saulnier, Guillaume. 2012. Ce que la fiction fait aux policiers. *Réception des médias et identités professionnelles: Travailler* 27: 17–36. [CrossRef]
Mauriello, Thomas P. 2020. *Public Speaking for Criminal Justice Professionals: A Manner of Speaking*, 1st ed. Boca Raton: CRC Press.
McDougall, Kirsty, Francis Nolan, and Toby Hudson. 2016. Telephone Transmission and Earwitnesses: Performance on Voice Parades Controlled for Voice Similarity. *Phonetica* 72: 257–72. [CrossRef]
McGehee, Frances. 1937. The reliability of the identification of the human voice. *The Journal of General Psychology* 17: 249–71. [CrossRef]
Morrison, Geoffrey S., and William C. Thompson. 2017. Assessing the admissibility of a new generation of forensic voice comparison testimony. *Columbia Science and Technology Law Review* 18: 326–434.
Morrison, Geoffrey S., Farhan H. Sahito, Gaëlle Jardine, Djordje Djokic, Sophie Clavet, Sabine Berghs, and Caroline Goemans Dorny. 2016. INTERPOL survey of the use of speaker identification by law enforcement agencies. *Forensic Science International* 263: 92–100. [CrossRef] [PubMed]
Nuance Communications. 2015. [White Paper]. The Essential Guide to Voice Biometrics. Available online: https://www.nuance.com/content/dam/nuance/en_us/collateral/enterprise/white-paper/wp-the-essential-guide-to-voice-biometrics-en-us.pdf (accessed on 23 January 2024).
Rafter, Nicole. 2007. Crime, film and criminology: Recent sex-crime movies. *Theoretical Criminology* 11: 403–20. [CrossRef]
Ratliff, Evan. 2022. Persona: The French Decepion [Audio podcast]. Pineapple Street Studios—Wondery. Available online: https://wondery.com/shows/persona/ (accessed on 23 January 2024).
Ribeiro, Gianni, Jason M. Tangen, and Blake M. McKimmie. 2019. Beliefs about error rates and human judgment in forensic science. *Forensic Science International* 297: 138–47. [CrossRef] [PubMed]
San Segundo, Eugenia, and Hermann Künzel. 2015. Automatic speaker recognition of spanish siblings: (Monozygotic and dizygotic) twins and non-twin brothers. *Loquens* 2: e021. [CrossRef]
Sidtis, Diana, and Jody Kreiman. 2012. In the Beginning Was the Familiar Voice: Personally Familiar Voices in the Evolutionary and Contemporary Biology of Communication. *Integrative Psychological and Behavioral Science* 46: 146–59. [CrossRef] [PubMed]
Smith, Peter Andrey. 2023. Can We Identify a Person from Their Voice? Digital Voiceprinting May Not Be Ready for the Courts. *IEEE Spectrum*, April 15.
Solan, Lawrence M., and Peter M. Tiersma. 2003. Hearing Voices: Speaker Identification in Court. *Hastings Law Journal* 54: 373–435.
Stevenage, Sarah V. 2018. Drawing a distinction between familiar and unfamiliar voice processing: A review of neuropsychological, clinical and empirical findings. *Neuropsychologia* 116: 162–78. [CrossRef] [PubMed]
Trainum, James L. 2019. The CSI effect on cold case investigations. *Forensic Science International* 301: 455–60. [CrossRef] [PubMed]
Villez, Barbara. 2005. *Séries télé, visions de la justice*, 1st ed. Paris: Presses universitaires de France.
Villez, Barbara. 2014. *Law and Order. New York Police Judiciaire. La Justice en Prime Time*. Paris: Presses Universitaires de France. Available online: https://www.cairn.info/law-and-order-new-york-police-judiciaire--9782130594239.htm (accessed on 23 January 2024).
Watt, Dominic, and Georgina Brown. 2020. Forensic phonetics and automatic speaker recognition. In *The Routledge Handbook of Forensic Linguistics*, 2nd ed. Edited by Malcolm Coulthard, Alison May and Rui Sousa-Silva. London: Routledge, pp. 400–15. [CrossRef]
Watt, Dominic, Peter S. Harrison, and Lily Cabot-King. 2020. Who owns your voice? Linguistic and legal perspectives on the relationship between vocal distinctiveness and the rights of the individual speaker. *International Journal of Speech Language and the Law* 26: 137–80. [CrossRef]
Yarmey, A. Daniel, A. Linda Yarmey, and Meagan J. Yarmey. 1994. Face and voice identifications in showups and lineups. *Applied Cognitive Psychology* 8: 453–64. [CrossRef]
Zuluaga-Gomez, Juan, Sara Ahmed, Danielius Visockas, and Cem Subakan. 2023. CommonAccent: Exploring Large Acoustic Pretrained Models for Accent Classification Based on Common Voice. *Interspeech* 2023: 5291–95. [CrossRef]

languages

MDPI

Article

Distinguishing Sellers Reported as Scammers on Online Illicit Markets Using Their Language Traces

Clara Degeneve [1,*], Julien Longhi [2] and Quentin Rossy [1]

[1] École des Sciences Criminelles, University of Lausanne, 1015 Lausanne, Switzerland; quentin.rossy@unil.ch
[2] AGORA Laboratory EA 7392, CY Cergy Paris University, 95000 Cergy, France; julien.longhi@cyu.fr
* Correspondence: clara.degeneve@unil.ch

Abstract: Fraud exists on both legitimate e-commerce platforms and illicit dark web marketplaces, impacting both environments. Detecting fraudulent vendors proves challenging, despite clients' reporting scams to platform administrators and specialised forums. This study introduces a method to differentiate sellers reported as scammers from others by analysing linguistic patterns in their textual traces collected from three distinct cryptomarkets (White House Market, DarkMarket, and Empire Market). It distinguished between potential scammers and reputable sellers based on claims made by Dread forum users. Vendor profiles and product descriptions were then subjected to textometric analysis for raw text and N-gram analysis for pre-processed text. Textual statistics showed no significant differences between profile descriptions and ads, which suggests the need to combine language traces with transactional traces. Textometric indicators, however, were useful in identifying unique ads in which potential scammers used longer, detailed descriptions, including purchase rules and refund policies, to build trust. These indicators aided in choosing relevant documents for qualitative analysis. A pronounced, albeit modest, emphasis on language related to 'Quality and Price', 'Problem Resolution, Communicationand Trust', and 'Shipping' was observed. This supports the hypothesis that scammers may more frequently provide details about transactions and delivery issues. Selective scamming and exit scams may explain the results. Consequently, an analysis of the temporal trajectory of vendors that sheds light on the developmental patterns of their profiles up until their recognition as scammers can be envisaged.

Keywords: cryptomarket; fraud; scammer; language trace; forensic linguistic

Citation: Degeneve, Clara, Julien Longhi, and Quentin Rossy. 2024. Distinguishing Sellers Reported as Scammers on Online Illicit Markets Using Their Language Traces. *Languages* 9: 235. https://doi.org/10.3390/languages9070235

Academic Editor: Alan Garnham

Received: 15 November 2023
Revised: 13 June 2024
Accepted: 19 June 2024
Published: 28 June 2024

1. Introduction

Illicit markets, like legal markets, have been transformed by the virtual environment, which has altered promotional strategies, sales processes, and the sharing of evaluations between buyers and sellers. Sellers of illicit products and services employ multiple approaches to promote their products to potential customers. Online spaces used for selling illicit products and services can be classified into two main categories. The first is that of dedicated sales sites created on the web in the form of online stores associated with unique domain names and with contents managed by the spaces' administrators. The other category is that of collaborative platforms: online communities that are shared environments in which sellers publish their shops or their ads in a pre-existing convergence setting. Sellers may focus on lawful commerce or the trade of illicit goods and services on specialised forums and cryptomarkets present on the dark web. As Martin (2013) explains, cryptomarkets are digital marketplaces hosted on the dark web that facilitate transactions primarily using cryptocurrencies as the medium of exchange. These marketplaces promote user privacy and anonymity, using a combination of encryption and routing techniques to obfuscate both the identities of the participants and their financial transactions. These markets were initially used to sell drugs but have since diversified into selling many types

of products, such as false identity documents and credit cards, as well as illegal services such as hacking services.

The online activities of sellers and buyers on these platforms leave digital traces that can be exploited to study illicit markets (Rossy and Décary-Hétu 2017). They offer a wealth of information relevant to addressing cross-cutting questions such as 'What substances are available?', 'What is the market volume?', 'What are the sellers' revenues?', and 'How is the market structured?' Regular monitoring of these spaces broadly allows for trend tracking and for considering the use of these traces as a monitoring indicator.

Such analyses, however, require the ability to assess the relevance of the information available on these platforms. Among the validity challenges of these indicators is the presence of fraudsters, who can skew the analyses with sales behaviours that do not match those of genuine sellers. A scam in this case can be described as (Jacquart et al. 2021, p. 410) when 'a person is interested in an item for sale on a classified ad site, contacts the advertiser, then pays the negotiated amount to the advertiser, but never receives the item'. Within cryptomarkets, the identification of fraudulent sellers can be facilitated in two primary ways: directly by the platform's administrators or through discussion forums where customers can report dubious activities and alert fellow users (Morselli et al. 2017).

However, can fraudsters be detected by analysing the traces they leave on platforms? Does their sales behaviour differ from that of genuine sellers? This article proposes an approach based on the forensic analysis of language traces left by sellers in their seller profiles and in disseminating their ads. The main hypothesis of this work is that it is possible to distinguish reported scammers from legitimate sellers on cryptomarkets using language traces (Renaut et al. 2017). A language trace can be defined as the remnant of an action that alters the environment (Ribaux 2023; Roux et al. 2022), which is the writing of an illegal or litigious text by an author and which has an informative potential not only for its source but also for the illicit activity itself (Degeneve et al. 2022).

2. Previous Research

2.1. Digital Traces Left by Trust Mechanisms in Online Trade

2.1.1. Deceptive Practices in Cryptomarkets

Most research on the detection of deception, which is a different deviant behaviour from fraud, has focused on the distinction between false and truthful statements, even if both may be considered deceptive activities. Research on the production and detection of deception typically concentrates on individuals, particularly on how they convey lies and the extent to which others can identify those lies (Markowitz et al. 2023). Scams are 'acts carried out deliberately to enhance one's gains at the cost of others' (Christin 2013, p. 241). Deception is used for obtaining illegal financial gain with 'the deliberate intent to deceive with the promise of goods, services, or other financial benefits that in fact do not exist or that were never intended to be provided' (Titus et al. 1995, p. 54).

Markowitz et al. (2023) emphasise that the context of a deception behaviour is pivotal, as outlined in their extended Contextual Organization of Language and Deception (COLD) framework. They argue that traditional deception studies often overlook the nuanced interplay between context and communicative behaviour, leading to inconsistent findings across different settings. By integrating three context dimensions—individual differences, situational opportunities for deception, and interpersonal characteristics—Markowitz et al. (2023) enhance the COLD model's applicability to forensic settings. Their study underscores the complexity of deception, suggesting that both the psychological dynamics of the deceiver and the specificity of the communicative environment significantly influence the effectiveness of deception detection strategies. This holistic approach advocates for a more nuanced consideration of context. As Hancock et al. (2004) note, deception is influenced by communication technologies (e.g., email, instant messaging, and telephone systems). We thus describe here the deception schemes underlying fraud in cryptomarket environments.

Morselli et al. (2017) conducted a study of the management of disputes and violent incidents in the drug segment of cryptomarkets. Drawing on Christin's (2013) observation that a vast majority (95%) of feedback is positive, the researchers explored the phenomenon of scamming. Further, the study agreed with Tzanetakis et al. (2016) in pointing out a tactic employed by some sellers of posting derogatory remarks to tarnish their competitors. The researchers gathered scam-related data from 10 cryptomarkets and a widely visited forum. The main areas of contention they discerned were the non-delivery of goods, lack of communication, and inferior product quality. Qualitative analysis of forums revealed that sellers accused of scamming are occasionally defended by peers, who urge aggrieved buyers to exhibit patience. The inherent design and framework of cryptomarkets aim to curb, if not eradicate, scam incidents.

In a 2020 study, Bancroft et al. (2020) analysed the drug-centric discussion forum PFM, focusing on how reputation is quantified through a 'Karma score' (a form of feedback given by fellow users) and various status indicators. The researchers outlined different manifestations of fraud within this environment. One notable tactic is the 'exit scam', wherein previously trustworthy sellers abruptly decide to deceive buyers. Additionally, Bancroft et al. (2020) highlight the phenomenon of 'selective scamming' (p. 9), in which a seller deliberately fails to dispatch a specific order and retains the payment yet remains undetected by fulfilling all other orders as expected.

The issue of fraudulent activities on cryptomarkets seems to mirror its significance in legal markets. Ineffective administration by market overseers has been identified as a root cause of such deceitful activities (Christin 2013). Many researchers have emphasised the pivotal role of discussion forums as platforms on which users share insights and highlight potential scams. However, the methodologies for identifying potential scammers can diverge.

2.1.2. The Informative Value of Feedback

Feedback is a pivotal element underpinning the concept of trust within online marketplaces. The essence of trust for sellers revolves around expanding their customer base. Feedback mechanisms on digital marketplaces, including cryptomarkets, provide buyers with tools to express their satisfaction or dissatisfaction with a transaction. As Przepiorka et al. (2017) explain, feedback mechanisms play a crucial role in shaping a seller's reputation, since they can directly impact the trajectory of subsequent transactions. Such mechanisms usually encompass two main features: rating systems and commentaries. Rating systems can range from simple thumbs up/down and star ratings to complex scorecards that evaluate different aspects of a transaction, such as product quality, shipping speed, and communication efficacy. Commentaries allow for qualitative feedback, enabling buyers to elaborate on their experiences, highlight specific strengths or concerns, and provide context for their numerical ratings. Hypothetically, the analysis of customer feedback on sales platforms or dedicated forums could aid in detecting a fraudster.

Pavlou and Dimoka (2006) delve into two aspects governing marketplace actors: benevolence and credibility. While benevolence pertains to a seller's genuine goodwill and positive intent towards buyers, credibility revolves around the seller's competency and reliability in fulfilling promises (p. 383). Pavlou and Dimoka analysed 10,000 eBay feedback comments to infer sellers' intentions based on buyers' perceptions. The data came from '1665 completed auctions for 10 distinct products (iPod, n = 512; movie DVD, n = 341; music CD, n = 312; Palm Pilot, n = 138; digital camera, n = 110; camcorder, n = 92; DVD player, n = 84; monitor, n = 76) during May of 2005' (p. 400). The analysis revealed occasional discrepancies between the magnitude of the rating and the sentiments of pronounced comments (especially those using terms like 'abominable' or 'absolutely fabulous'). The subsequent phase involved disseminating a survey to purchasers via email to glean demographic information, with a total of 420 responses received. In the third phase, a content analysis was performed on feedback comments, limiting the scope to the first 25 comments for each seller. This analysis was conducted by a team of three individuals

who categorised comments into five distinct groups. The 25 comments evaluated for each of the 420 sellers revealed that many buyers consult the comments prior to making a purchase. Distinct patterns emerged between benevolence and credibility in the comments, though they are not mutually exclusive. Overall, these comments offer deeper insights into a seller's reputation than mere ratings do. Notably, those categorised as 'ordinary' predominantly carried a positive sentiment.

In their research on Silk Road 1.0, Przepiorka et al. (2017) derived data specific to drug transactions. They utilised several key indicators: total ratings received, product pricing, pace of sales, and feedback count per product. Analysing a dataset of 3153 products, they observed that a staggering 95.8% of ratings were perfect 5/5 scores. The researchers subsequently designed a model to assess how fluctuations in an individual seller's reputation, when juxtaposed against their mean reputation, would influence sales dynamics. Their findings underscored a direct correlation: sellers with sterling reputations were inclined to elevate their pricing and expedite sales, and most interestingly, the negative feedback bore a heftier impact than its positive counterpart.

Tzanetakis et al. (2016) draw a parallel between traditional narcotics markets and those operating on cryptomarkets. Street data, both qualitative and quantitative in nature, were sourced from a German project. The research focused on the motivations behind individuals' involvement in drug trafficking, narrowing down the sample to 32 individuals who were primarily driven by monetary incentives. Data from Agora included profiles, feedback, and forum discussions derived from four active sellers, two high-rated and two low-rated. Within these groups, one seller offered multiple drug varieties while the other was limited to one or two. Sales figures revealed three sellers with transaction counts ranging from 200 to 500, and one with transactions between 2000 and 3000. Trust dynamics differed markedly between traditional and cryptomarkets, with the former emphasising safety. On cryptomarkets, trust determinants are intrinsically tied to platform-provided metrics, such as sales volumes, ratings, and general seller information. Ratings predominantly ranged from four to five. Dark web forum interactions substantially foster trust, but some sellers exploit this by leaving negative feedback to tarnish their competitors.

In their recent study, van Deursen (2021) undertook a qualitative examination of feedback on the AlphaBay cryptomarket to investigate the repercussions of the polarity of forum posts on the vendors' sales and pricing. They classified feedback as 'positive', 'neutral', or 'negative' and found that it was often supplemented with descriptive comments. The platform also features an embedded forum. In total, 1655 articles, inclusive of seller details, feedback, and relevant forum content, were gathered. The determination of polarity was achieved through sentiment analysis using a Random Forest algorithm. The findings suggest that positive textual feedback carries more weight than ratings in influencing a seller's market presence. Intriguingly, and contrary to expectations, negative feedback has a beneficial impact on sales and a more pronounced influence on pricing than positive feedback. Among forum comments, positive ones were found to be the most impactful.

Several key observations drawn from previous research informed the direction of our study. First, high feedback scores, though ostensibly indicative of a trustworthy seller, can be misleading. The overwhelmingly positive ratings across various platforms underscore the potential pitfalls of using this metric in isolation to gauge the legitimacy of sellers. Second, forums stand out as crucial platforms where customers exchange insights and raise alarms about dubious actors. The interactive nature of these platforms, coupled with users' collective experiences, often results in a more accurate depiction of a seller's reputation. Nevertheless, the question remains: is it possible to detect a scammer based on their activity within a cryptomarket?

2.1.3. Leveraging Language Traces to Unveil Fraud through Computational Linguistic Analysis

We hypothesise that fraud in online settings—as a particular type of deceptive behaviour like lies, fake news, or rumours—can be detected based on the analysis of language

traces that include deceptive linguistic cues. Language traces are pieces of information used to change people's cognition or beliefs (Addawood et al. 2019). Addawood et al. (2019) explored the linguistic indicators of deception used by political trolls on social media to mislead the audience about their true intentions. They highlight the vulnerability of content-focused social media platforms to such tactics due to their reliance on asynchronous, text-based communication. Through their analysis, they identified key linguistic cues employed by trolls, such as persuasive language, simpler and less specific language, and a higher frequency of hashtags, tweets, and retweets. Despite the challenges of detecting trolls due to their rarity and the resulting imbalance in classification tasks, Addawood et al. note that troll accounts often use fewer nouns and post shorter tweets. They conclude that a simple algorithm could mistakenly classify most accounts as non-trolls with high accuracy but low recall.

Computational linguistics and machine learning experts have also ventured into fraud detection, employing diverse methods to tackle the issue.

Gibbons and Turell (2008) present Egginton's examination of an advanced fee fraud's linguistic framework. Drawing from a spam email he obtained, Egginton undertook discourse analysis to discern tactics for bolstering the potential victim's trust based on that specific spam email. These include specificity in the email's subject and main content, highlighting the situation's urgency, and using complex technical jargon.

Ott et al. (2011) present a comparative approach for detecting fake reviews on Trip Advisor. They compare the results obtained by a panel of three human judges with those of an automated detection algorithm. While the human panel achieved approximately 60% performance in detecting fake reviews, the automated approach reached around 89% performance. Analysing unigrams and bigrams of words proved to be one of the most effective methods (Ott et al. 2011).

Vidros et al. (2017) explored fraudulent online job advertisements. Utilising a corpus of 450 deceptive and 450 genuine ads from an online platform, they applied the bag-of-words model for vectorisation. Subsequently, they trained six distinct classifiers ('ZeroR, OneR, Naive's Bayes, J48 decision trees, random forest and logistic regression' p. 9). The Random Forest algorithm proved the most accurate, with a 91% precision rate. Through a linguistic and contextual analysis, the team determined that deceitful ads were generally shorter, lacked details about job requirements and perks, and exhibited certain keyword patterns such as 'home' (implying 'work from home'). Finally, binary analysis of the corpus showed some characteristics that can help with the distinction: '(a) opportunistic career pages usually do not have a corporate logo; (b) scammers omit adding screening questions; (c) usually mention salary information even in their title to lure candidates; (d) skip designated job attributes (i.e., industry, function, candidate's education level, and experience level) used for jobs board categorisation; (e) prompt defrauded candidates to apply in external websites, bypassing the ATS pipeline; (f) or force them to send their resumes to their personal email addresses directly and (g) address lower educational level' (p. 12).

Jakupov et al. (2022) employed topic modelling to discern deceptive opinion spam among reviews on Trip Advisor, amassing a dataset of 6977 reviews. Utilising the BERTopic module in Python, their methodology effectively identified lexical indicators of deceit within the texts ('Max size of N-gram dictionary: total number of rows in the n-gram dictionary; Rho parameter: prior probability for the sparsity of topic distributions; Alpha parameter: prior probability for the sparsity of topic weights per document; Size of the batch: number of rows processed in chunks; Initial value of iterations used in learning update schedule: learning rates start value, set to 0 in all the experiments; power applied to the iteration during updates: learning stepsize; N-grams: the maximum size of the sequences generated during hashing') (Jakupov et al. 2022, p. 7).

Junger et al. (2023) address the gap in understanding how fraud victims and near-victims recognise deception in real-world scenarios. The researchers analysed responses from a victimisation survey to elucidate effective deception detection strategies used

by individuals exposed to various types of online fraud. They revealed that 69% of near-victims had recognised fraud through their existing knowledge of fraud tactics and warning signs, such as inconsistencies or errors in the fraudsters' communications. Other detection strategies included distrust, adherence to personal security rules, and seeking additional information. Victims and near-victims often cited past experiences and increased awareness from media exposure as factors enhancing their ability to spot fraud. Junger et al. emphasise that knowledge of fraud and proactive information-seeking behaviours are the most effective defences against fraud victimisation.

Research has hinted at the potential of language analysis in discerning deceitful behaviours, as in the studies of deceptive opinion spam and fraudulent listings described above. However, this approach remains largely uncharted when it comes to analysing data from cryptomarkets. Our research seeks to bridge this knowledge gap. By employing a forensic analysis of language traces left by sellers in both their profiles and their advertisements, we developed a methodological approach that harnesses the nuances of language to uncover deceptive patterns. We designed a refined toolset for discerning genuine sellers from reported scammers. The fusion of linguistics and forensic science with the rich data of cryptomarkets could provide an innovative method for vetting the authenticity of sellers, fortifying the integrity of analyses derived from these platforms.

3. Materials and Methods

3.1. Datasets

The data used for this research were collected from three cryptomarkets: DarkMarket, Empire Market, and White House Market (refer to Table 1), which were the major cryptomarkets on the dark web at the time of the study. Initiated in January 2018 on the dark web, Empire Market stepped in to fill the gap created by the mid-2017 closure of AlphaBay Market and quickly rose to prominence, remaining one of the largest dark web markets until August 2020. DarkMarket then rose to become the largest illicit dark web marketplace of its time. It was shut down in January 2021 by an international task force coordinated by Europol. Collection for the White House Market was performed over the period from April 2020 to March 2021. This cryptomarket opened in February 2019 and was closed by its creators in October 2021.

Table 1. Data collected from DarkMarket, Empire Market, and White House Market.

	N of Unique Ads	N of Unique Description	Crawling Period	Nb Crawls
EM	87′543	61′346	2020.06–2020.08	8
DM	88′640	45′432	2020.07–2021.01	17
WHM	83′524	56′739	2020.04–2021.03	30
Total	259′707	163′517		

With the objective of discerning reported scammers within the pool of sellers, our research utilised the Dread forum, as guided by previous scholarly findings. Dread is a dark web forum in operation since 2018.[1] It is structured, similar to Reddit, by threads dealing with different subjects, and users can exchange ideas, create posts, and reply to each other within the same post. In particular, there are threads dedicated to cryptomarkets where buyers and sellers can interact; these threads were collected, as they are also used by users to report fraud.

We methodically retrieved threads pertaining to the trio of cryptomarkets (refer to Table 2) and subsequently applied manual inspection of the posts to capture usernames that were frequently reported as fraudulent by peers.

From the cryptomarket data, we extracted profiles and advertisements corresponding to the identified usernames (for examples of profile and product description, see Appendix A, Figures A1 and A2). Based on the list of usernames, derived as mentioned

earlier, vendors were categorised as potential scammers. To respect privacy, no vendor names have been disseminated, and no other identifying information was used during the study. All the analyses were based on the text, and the results are presented such that no link can be established with the virtual identity of the sellers.

Table 2. Number of posts containing 'scam' for every cryptomarket thread on Dread forum.

Thread	Number of Posts Containing 'Scam'
White House Market	261
Empire Market	1967
DarkMarket	678
Total	2906

3.2. The Choice of a Computational Linguistic Approach

From a forensic perspective, computational linguistics offers objectivity, reproducibility, and accuracy (Juola et al. 2006). Nevertheless, as Fobbe (2020) points out, computational research in forensic authorship attribution lacks theoretical engagement. The assumption that language differences inherently indicate different authors or types of authors based on detectable characteristics should rely on robust methods. Surface structure features dominate studies, yet they fail to capture deeper stylistic nuances. The issue lies not in feature extraction but in the inadequacy of current style theories to explain how feature frequency correlates with individual authorship. Nevertheless, variation in the type of analysis we propose is not aimed at simply increasing the levels of analysis but at combining the analysis of different levels of language within a communicative conception of grammar and expression, as developed by Charaudeau (1992) in particular. We followed a 'method for decomposing the data into smaller chunks so that a larger set of variables can be used for the discriminant analysis' (Chaski 2005, p. 11) to fit with the forensic analysis criteria (Roux et al. 2022), and we also took a multi-level approach to linguistics, combining different markers that have in common the enhancement of certain communicative intentions. Juola (2021) has contrasted human and computational analysis in forensic science, noting the difficulty of establishing the validity and reliability of human-based analysis. Juola suggests prioritising objective features like shared vocabulary, word length, N-grams, common words, and punctuation. Longhi (2021) summarizes the tensions between the two approaches: while qualitative stylistic approaches can be seen as too subjective, 'much of this criticism comes from the United States, where the admissibility of expert evidence is determined in relation to the standards of the Daubert Criteria' (Wright 2014, p. 19). Thus, computational approaches are 'considered to be more objective, empirical, replicable, and ultimately more reliable than their stylistic counterparts', but they can hardly give information about theoretical aspects of linguistic variation.

3.3. Pretreatment

We employed the 'langdetect' Python module (accessible at https://pypi.org/project/langdetect/, accessed on 6 November 2023) for the task of language identification. We observed that a predominant proportion, 91.45% of genuine profiles and 91.20% of scam profiles, were in English. We opted not to translate the non-English texts into English for several compelling reasons. Firstly, translation inherently introduces alterations in stylistic elements, which could compromise the authenticity and integrity of the original text. Secondly, from a forensic perspective, modifying the original trace of a text through translation is not desirable, as it may obscure or alter linguistic traces pivotal to our analysis. Thirdly, stylometric comparisons across languages can be fraught with challenges. Factors such as sentence complexity and length can vary significantly between languages, making it difficult to draw accurate and consistent conclusions. Additionally, translation might inadvertently introduce translator biases, further complicating the authenticity and interpretation of the results. We deemed it crucial to preserve the original nuances

and subtleties of each text to ensure the reliability and robustness of our computational linguistic methods. Given that most of our sample consisted of English texts, we believed it reasonable to filter out non-English entries to maintain uniformity in the analyses.

All profiles and advertisement descriptions were filtered to retain only the unique texts for each seller on each of the cryptomarkets.

3.4. Analysis

We conducted a preliminary analysis of the number of distinct texts per vendor and per cryptomarket. Then, we subjected the corpus to a comprehensive examination utilising three distinct analytical methods: (1) textometric analysis to quantify text-based characteristics; (2) stylometric analysis, which includes syntax analysis to dissect sentence structures and linguistic patterns; and (3) N-gram analysis to discern overarching topics.

3.4.1. Textometric Analysis of Raw Texts

For the computational assessment of textual data, our study employed the 'textacy' Python module, which is accessible at https://textacy.readthedocs.io/en/latest/ (accessed on 6 November 2023). This tool enabled us to analyse textual statistics, including the number of characters, words, long words, unique words, and sentences. The number of uppercase letters was also calculated. The module also allowed for the evaluation of textual entropy, along with three readability and four diversity indices. The metrics used, along with their detailed definitions and references, are documented online at https://textacy.readthedocs.io/en/latest/api_reference/text_stats.html (accessed on 6 November 2023).

To assign grammatical labels to individual tokens within each corpus, we employed the Part of Speech (POS) Tagging features of the 'textacy' package. Subsequently, we computed the mean and median values of the analysed tags across all corpora per vendor.

Features such as the count of uppercase letters, readability scores, diversity indices, and POS tagging are instrumental in stylometric analysis to differentiate and compare authors or distinct bodies of text. These metrics are detailed alongside conventional textometric attributes, as they were derived from the raw text with minimal pre-processing that included only the elimination of return and tab characters. This choice allowed the exploration of textual features directly from unmodified text (i.e., language traces as they were left by writers).

3.4.2. Analysis of the N-Grams

The predominant words, bigrams, and trigrams of words in the lemmatised text were extracted, as proposed by Ott et al. (2011), through the large language model of the 'spacy' algorithm integrated into the 'textacy' module. This was preceded by a comprehensive pre-processing routine that standardised bullet points, quotation marks, and whitespace while excluding punctuation. The frequency analysis of those N-grams allowed for the identification of recurring topics.

3.4.3. Probabilistic Discrimination

In line with a recent paper on ChatGPT authorship discrimination (Bozza et al. 2023), we compared the use of N-grams between reported scammers and all other vendors using a probabilistic approach (Aitken and Taroni 2005; Taroni et al. 2022). This approach computes the ratio of the probability of occurrence of a particular form if the vendor is labelled as a 'scammer' divided by the probability of occurrence of the same form for all other vendors. This likelihood ratio (LR) score signifies the overuse of a form by one group over the other. A value exceeding one supports the hypothesis that the behaviour is more characteristic of the reported scammers over other vendors, while a value less than one indicates a preference for the alternative hypothesis, suggesting that the behaviour is more typical of all other vendors. For instance, an LR of two indicates that it is twice as probable to see the word if the text was written by a reported scammer. Before calculating the LRs, all forms used by less than 1% of vendors were filtered.

4. Results

4.1. Textometric Analysis of Raw Texts

The number of unique descriptions per seller profile over the collection period differed between reported scammers and other sellers (refer to Table 3). On DarkMarket and Empire Market, the average number of distinct profiles increased from 1.0 to 1.6 and from 1.2 to 1.8, respectively. The rise in the White House Market was less pronounced, from 1.6 to 1.9. The analysis of ad descriptions yielded a global ratio of 1.24 distinct texts per ad for those classified as scams, which is notably comparable to the ratio of 1.15 calculated for non-scam advertisements.

Table 3. Number of distinct profiles and ad descriptions for reported scammers and others on each cryptomarket.

	Reported Scammer N Profiles > N Distinct Descriptions	Others N Profiles > N Distinct Descriptions	Reported Scammer N Ads > N Distinct Descriptions	Others N Ads > N Distinct Descriptions
DarkMarket	33 > 52	809 > 813	717 > 906	33′465 > 39′645
Empire Market	83 > 153	804 > 936	2099 > 2537	43′536 > 48′767
White House Market	73 > 142	2036 > 3240	2632 > 3308	42′464 > 49′326
Total	189 > 347	3647 > 4925	5448 > 6751	119′465 > 137′272

The distributions of textual statistics showcased in Figures 1 and 2 fail to differentiate reported scammers from other sellers in terms of their profiles or their ad descriptions.

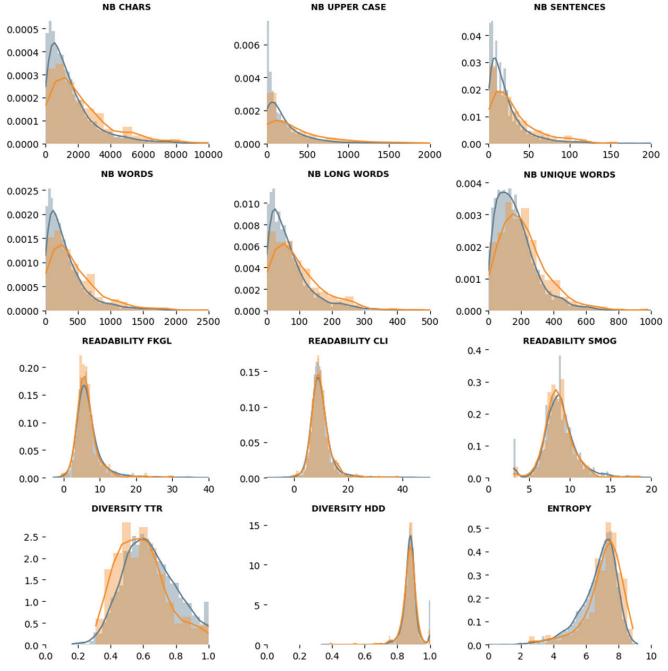


Figure 1. Textometric analysis of profile descriptions. The chart illustrates a comparative analysis with scam-related advertisements depicted in orange at the top, while other advertisements are represented in grey.

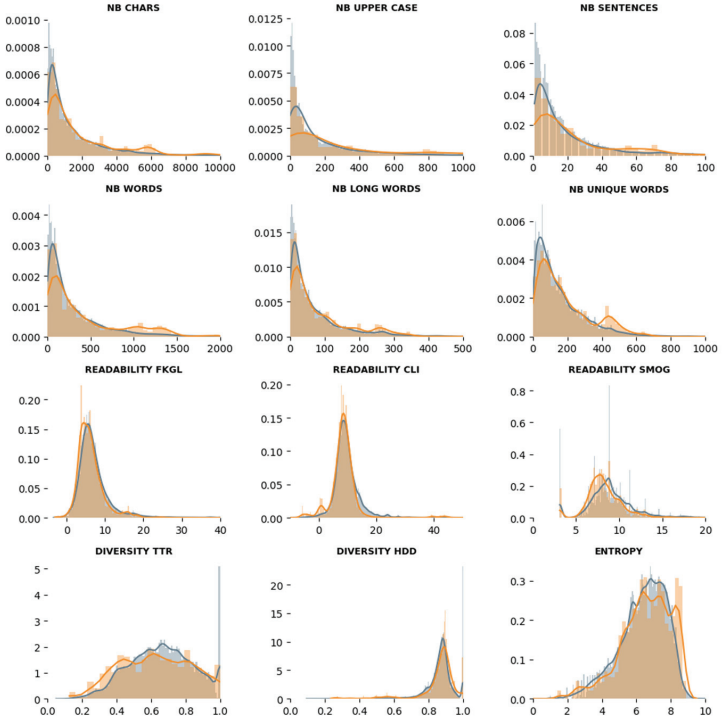


Figure 2. Textometric analysis of product descriptions. The chart illustrates a comparative analysis with scam-related advertisements depicted in orange at the top, while other advertisements are represented in grey.

However, the frequency distribution of word counts used by reported scammers, particularly for unique words, exhibits a notable peak ranging from 1000 to 1500 words and between 400 and 500 unique words which are also associated with high entropy (see Figure 2). Indeed, these texts include lengthy descriptions that incorporate details such as purchase rules, (non-)refund policies, delivery times, and other FAQ-related information. In some cases, the vendor's PGP key is provided within the message so that the buyer can verify the seller's identity through encrypted messaging. Overall, these messages seem to contain a substantial amount of information aimed at maximising buyer trust. They feature phrases like 'we never should SCAM you and we know how it's to get scammed'.

Additionally, a focused analysis was conducted at the peak of texts with low 'readability-cli' scores. These texts primarily contain repetitions of special characters such as '=', '-', '+', '*', '#', and a substantial number of emojis. Indeed, Coleman and Liau's (1975) algorithm relies on the number of characters instead of the number of syllables or words. By filtering these tokens, it was possible to identify types of advertisements that maintain a low-level readability score, which are very brief texts (3–10 words). Within this group, 'custom listing' (Soska and Christin 2015) and 'tip jar' listings were detected. Other texts contain lists of short sentences.

As shown in Figures 3 and 4, the POS-tagging analysis of the listings and profiles did not reveal any significant differences for sellers labelled as scammers. Principal component analysis (PCA) conducted with every available tag revealed that the data is overall inseparable (see Figures 3 and 4). Except for some outliers, PCA does not allow for a clear distinction between specific groups.

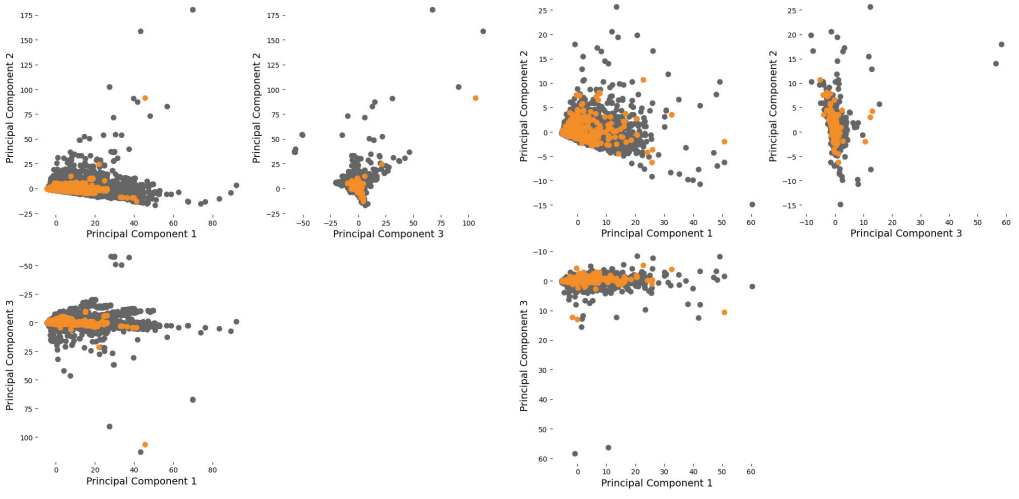


Figure 3. Result of the three-dimensional principal component analysis performed on the 66 POS-tags (listings on the left and profiles on the right). The chart illustrates a comparative analysis with scam-related advertisements depicted in orange at the top, while other advertisements are represented in grey.

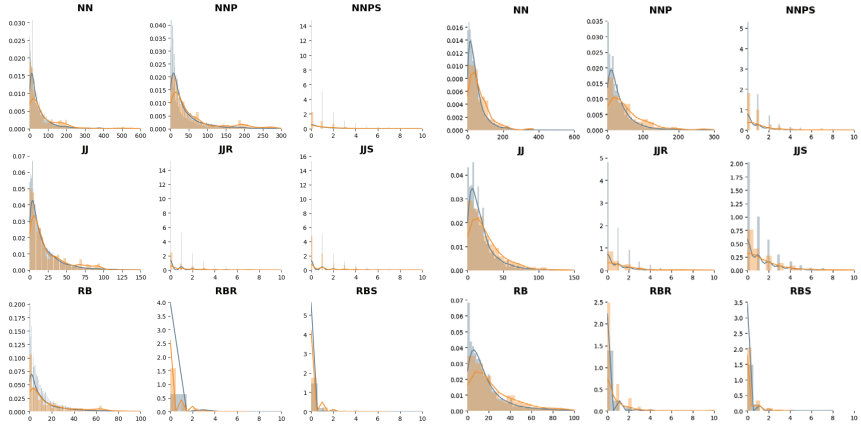


Figure 4. POS-tagging analysis of product descriptions (**left**) and profiles (**right**) with eight tags (NN—Noun, singular or mass, NNP—Proper noun, singular, NNPS—Proper noun, plural, JJ—Adjective, JJR—Adjective, comparative, JJS—Adjective, superlative, RB—Adverb, RBR—Adverb, comparative, RBS—Adverb, superlative). The chart illustrates a comparative analysis with scam-related advertisements depicted in orange at the top, while other advertisements are represented in grey.

4.2. Analysis of the N-Grams

Figure 5 illustrates that no N-gram distinctly stands out in differentiating reported scammers from other vendors. Indeed, an N-gram would be inherently discriminatory if it were positioned in the top left for words specific to them and in the bottom right for words specific to other vendors. Indeed, all the words used by more than 30% of reported scammers have a likelihood ratio ranging between 0.9 and 1.9. Therefore, no specific word seems to be discriminatory. Only the bigram 'high quality' appears to be relatively frequent, found in 50% of listings as opposed to 37% in those of other vendors.

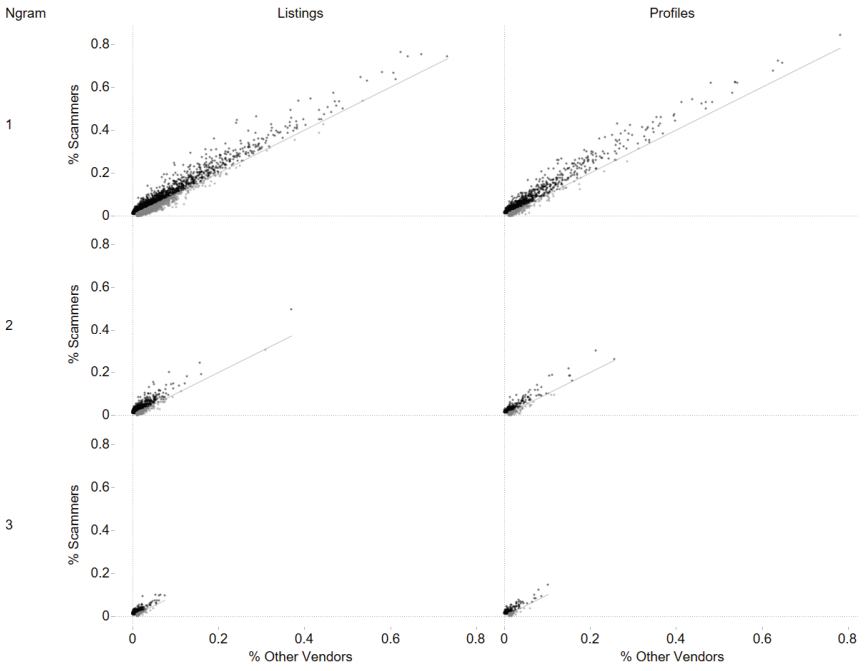


Figure 5. Global overview of the analysis of N-grams. Each point represents a specific N-gram plotted based on its usage frequency by reported scammers (Y-axis) and other vendors (X-axis).

4.2.1. Analysis of the Product's Descriptions

Numerous words and N-grams within the dataset pertain to marketed products (see Figures 6 and 7), while the remaining subset aligns with the lexical domain of sales. Figure 6 focuses on unigrams used by more than 30% of the reported scammers. Terms with the highest LRs are linked to the description of the product type, which is mainly related to drugs. Notably, most terms exhibit an LR falling within the range of 1.1 to 1.3, with none registering below 1. This implies that except for 'address' (LR = 0.9), all unigrams analysed in this context are slightly more prevalent among reported scammers. Most of these terms are in the lexical field of shipping and sales conditions.

Quality and price: The only recurrent bigram is 'high quality', which is employed by almost 50% of reported scammers; its LR of 1.4 indicates that it is only 1.4 times more prevalent among scammers than legitimate vendors. The trigram 'high uality product' is used similarly among both groups and might be a common marketing phrase used to build confidence in the product, regardless of the vendor's legitimacy. The percentage of reported scammers utilising other bigrams is notably low, except for 'lab test' and 'good quality' which are also related to the 'Quality' topic. Additionally, the words 'high dose', 'high purity', and 'high THC', which are grouped in the 'high' category, can also be integrated into the lexical field of product quality.

Problem resolution, communication, and trust: The forms 'reship policy', 'negative feedback', 'refund policy', and 'refund reship' that we decided to regroup with the negation forms 'doesn't' and 'don't' share a commonality in the context of what we called 'problem' (i.e., problem-resolution phrases). They are terms associated with policies, procedures, and customer service aspects, particularly in relation to the handling of disputes, returns, and customer satisfaction. They can be indicative of a seller's approach to handling issues like returns (reship and refund policy), customer complaints (negative feedback), and general terms of service or product guarantees. Communication-related trigrams such as 'question feel free', 'free to contact', and 'let we know' are also quite common. Lastly,

'term and condition' is also a common phrase used frequently in scams (6.5%) and other listings (6.4%). This could be because all vendors want to establish a sense of formality and legitimacy to increase trust.

Product-specific terms are mainly linked to drug names like '2cb pill', 'ketamine s' (linked to the trigrams 'ketamine s isomer' and 's isomer ketamine'), 'og kush', 'mg mdma', and 'xtc pill', which is not surprising given that these cryptomarkets are primarily used for drug sales.

Overall, while certain bigrams and trigrams are used more by reported scammers, many are also common in the overall corpus. This indicates the complexity of distinguishing them based solely on N-gram analysis. It suggests the need for more sophisticated methods or additional variables to accurately identify scam listings.

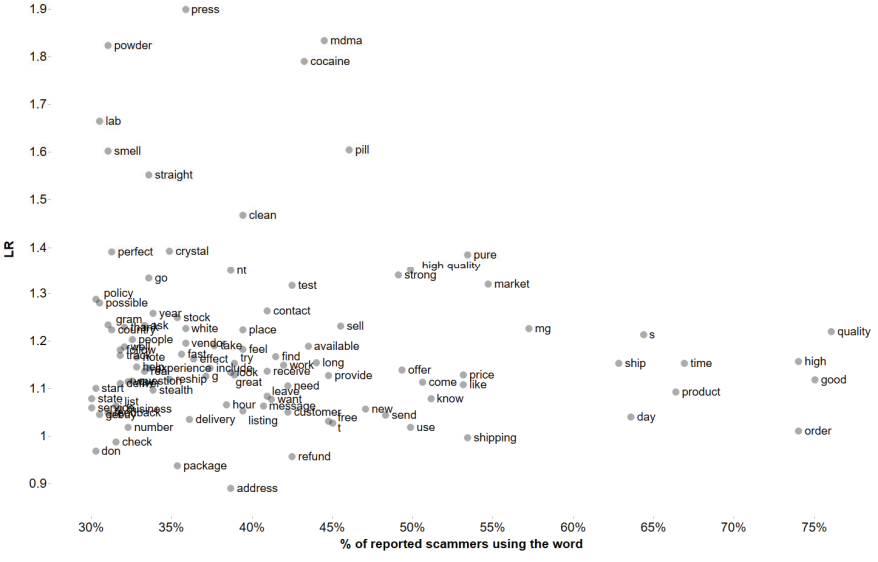


Figure 6. Likelihood ratio of unigrams in the corpus of product descriptions.

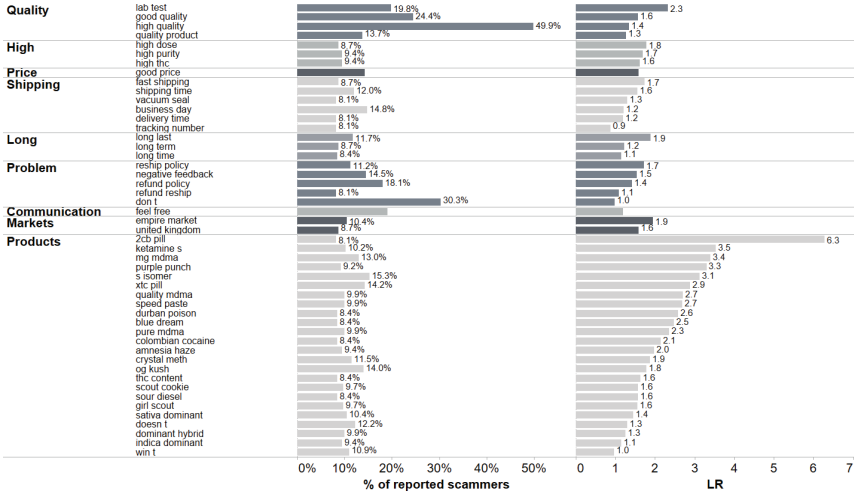


Figure 7. *Cont*.

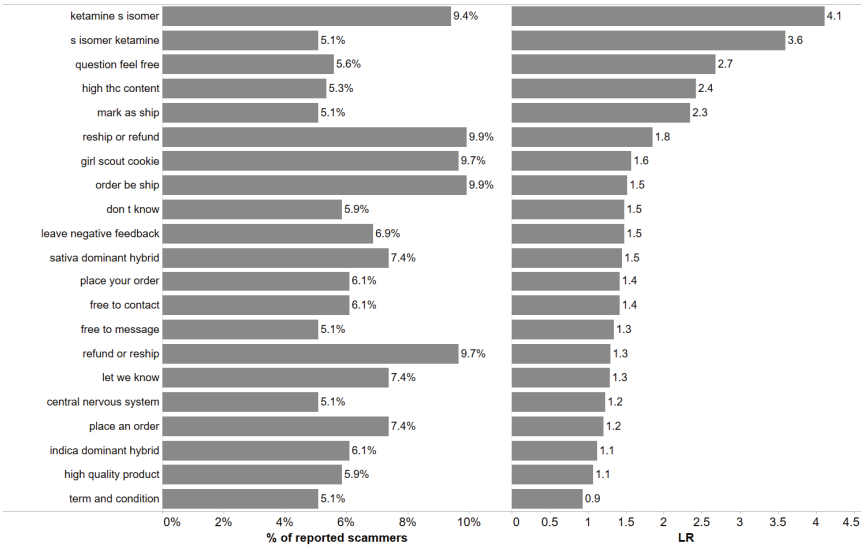


Figure 7. Bigram and trigram analysis of the listing corpus.

4.2.2. Analysis of the Profile Descriptions

This analysis offers insights into the linguistic patterns prevalent in the profile descriptions of reported scammers. The examination of single-word usages is detailed in Figure 8, while the analysis of bigrams and trigrams is delineated in Figure 9. These figures collectively reveal the verbal strategies scammers employ in their profiles.

The lexical domains of shipping and sales persist consistently (see Figure 8). Most unigrams are employed by 30–50% of reported scammers, with LRs between 1 and 1.3. This outcome closely mirrors that obtained from the listing corpus. Notably, the term 'order' is used by 84% of reported scammers but with an LR of 1.1, indicating that its usage is not significantly higher among scammers compared to legitimate vendors. It is worth noting that unigrams with the highest LR values (ranging from 1.55 to 1.65) are partially associated with shipment ('country', 'way', and 'fast').

Quality and price: The high incidence of 'high quality' in scam listings indicates that reported scammers recurrently advertise the quality of their products. However, like the terms 'good quality' and 'good price', it is quite common in the overall corpus, potentially diluting its effectiveness as a distinguishing feature.

Problem resolution, communication, and trust: The term 'customer support' has a higher likelihood ratio, which indicates it is six times more common in reported scammers' profiles. This suggests that scammers may prioritise establishing a facade of trust and support to attract and reassure potential customers. The higher occurrence of 'reship policy' and 'refund policy' in scam listings could be an effort to appear as though they provide customer protection and service, which could lower the perceived risk. Notably, no single trigram is overwhelmingly used, as even the most frequently trigram used by reported scammers ('refund or reship') occurs in only 14.7% of listings compared to 10.2% of non-scam listings. This suggests a subtle overlap in the language used by both groups of sellers. Trigrams such as 'reship or refund' and 'leave negative feedback' are quite common in scammers' profiles. However, the difference is not stark, with 'leave negative feedback' appearing in 10% of scam listings versus 6.9% of clean listings. 'Feel free', 'free to contact', and 'customer service' are common in both scam and other profiles. The data show that reported scammers do not use a drastically distinct set of bigrams and trigrams compared to other vendors.

Shipping and time: Seller profiles contain more information than their listings about shipping times: 'shipping time' (18.2%), 'business day' (18.2%), and 'delivery time' (12.9%). This is probably in part due to the period of the collection, which was during COVID-19. More globally reported scammers frequently use shipping-related terms. Forms like 'track order', 'post office', 'po box', 'postal code', and 'address format' are associated with the logistics of sending and receiving goods, which might emphasise the need to guide and reassure buyers of the transaction process. Overall, scammers might strategically use shipping-related forms to build credibility and simulate reliability in the delivery process. Nevertheless, these terms relating to standard business operations and logistics are less distinct and may not be reliable indicators on their own.

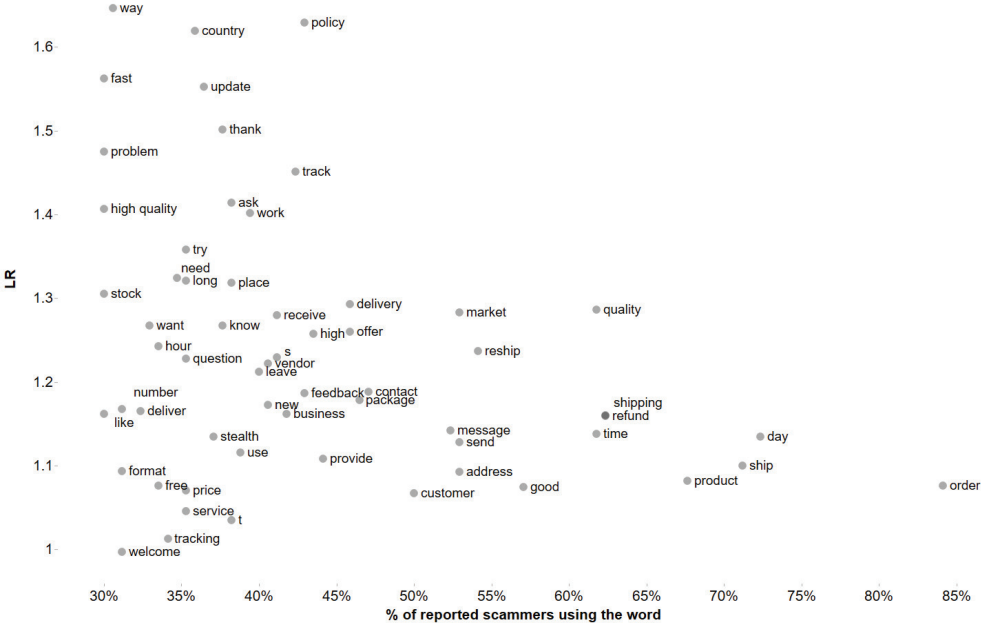


Figure 8. Likelihood ratio of unigrams on the corpus of profiles.

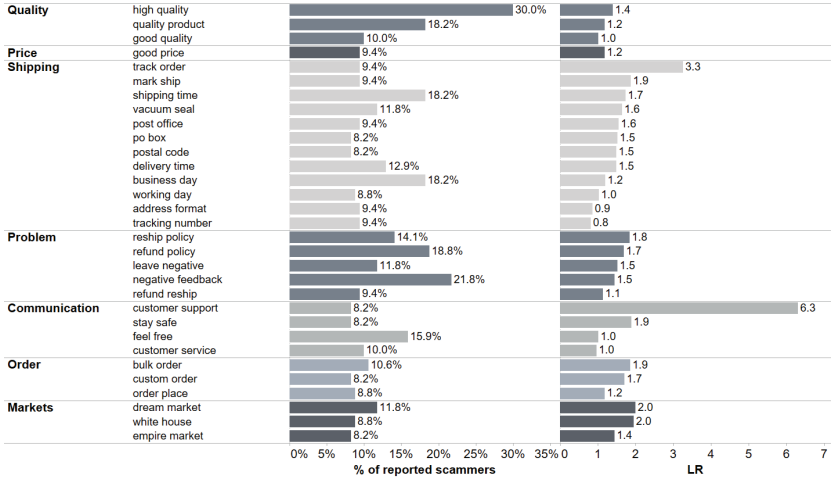


Figure 9. *Cont.*

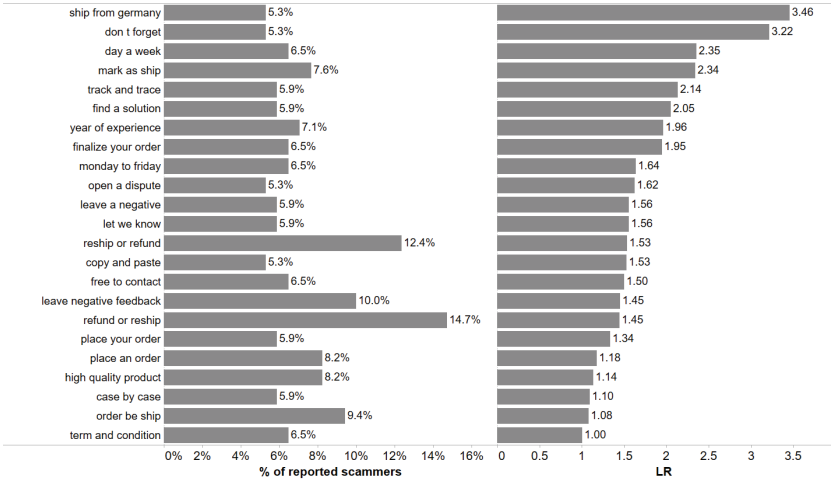


Figure 9. Bigram and trigram analysis of the profile corpus.

Order types: Mentions of 'bulk order' and 'custom order' in scam listings highlight sellers offering deals that appear more personalised or financially beneficial.

Marketplace names: References to specific markets, such as 'dream market', 'white house', and 'empire market', are used by sellers to link their current profiles with accounts on other marketplaces.

In summary, reported scammers seem to use language that aims to build trust, emphasise the shipping process, and stress the quality of their products to entice potential buyers. While some terms are more prevalent in scam listings, many are also commonly used by all vendors, which presents a challenge for distinguishing between the two based on language alone.

5. Discussion

What really is a scammer? The results obtained in this study might be explained by the fundamental definition of what is referred to as a 'scammer' on cryptomarkets. The initial recognition of scammers, leading to the construction of the corpus, was based on self-regulation. The demarcation between scammers and legitimate vendors is contingent upon user-reported allegations on the Dread forum and thus lacks assurance that the sellers accused of perpetrating scams are unequivocally scammers. This reflects a broader issue within online marketplaces, where accusations can be both a reflection of true misconduct and a tactic in competitive sabotage, as noted in studies of online behaviour and marketplace dynamics (Soska and Christin 2015; Morselli et al. 2017).

Furthermore, the taxonomy used for categorising these vendors fails to accommodate the conceptual distinctions inherent in selective scamming and exit scamming (Morselli et al. 2017; Bancroft et al. 2020; Décary-Hétu et al. 2018; Morselli et al. 2017). Selective and exit scammer vendors engage in fraudulent activities sporadically, maintaining conventional vending practices for the remainder of their operations. Consequently, their indistinguishability from standard vendors can be attributed to this phenomenon. In the context of selective scamming, vendors conduct most of their transactions legitimately, interspersing them with occasional deceptive practices. In contrast, exit scammers perpetuate a facade of normalcy in their transactions until a strategic juncture at which they abscond with funds and vanish from the platform. Such mixed behaviour is also described by Markowitz (2023), who notes the nuanced dynamics of deception within communication, challenging the traditional binary categorisation of statements as purely false or truthful. This indicates that the embedding of deceptive elements into truthful content is more complex than previously thought, and it calls for a deeper understanding of how deceptive elements

are interwoven into communications. The observation is also underscored by Hauch et al. (2015), who found consistent, albeit small, correlations between specific language patterns and deception.

Consequently, our analysis of textual statistics, including the numbers of characters, words, long words, unique words, and sentences, as well as the syntactic tags, exhibited no statistically significant divergences between the profile descriptions and advertisements. This suggests that textual analysis alone may not be sufficient for scam detection, which indicates a need for multimodal approaches that integrate other indicators such as the number of sales, the number of won and lost disputes, scores, or the quantities of positive and negative feedback. Textometric indicators, however, helped to identify peculiar ads in which sellers created longer descriptions with more details such as purchase rules, (non-)refund policies, and other FAQ-related information to increase trust. These indicators thus support the selection of pertinent documents on which a qualitative analysis can be focused.

The analysis of N-grams in both the listings and the profiles in our corpus revealed only minor differences between reported scammers and all other vendors regarding their content. It is also worth noting the relatively low percentages of those N-grams across the corpus, which indicates that there is not a 'silver bullet' bigram or trigram that clearly flags a listing as a scam. This could make it difficult for automatic detection systems to rely solely on these N-grams without a significant number of false positives. Nonetheless, one outcome of the N-gram analysis supports the differentiation: we observed an overuse of the lexicon pertaining to the 'Quality and Price', 'Problem Resolution, Communication, and Trust' and 'Shipping' topics. This suggests that scammers might offer more detailed information about transactions and delivery, potentially alleviating customer concerns. In these environments, where physical goods cannot be inspected before purchase, the trustworthiness and reputation of a seller are paramount. This pattern could indicate a strategic overcompensation, aligning with the deception strategies described in the literature, in which scammers create narratives to build trust (Button et al. 2014; Rossy and Ribaux 2020).

Additionally, the hypothesis that vendors adapt their descriptions in response to accusations of scamming remains a plausible explanation for the linguistic patterns we observed. In the context of cryptomarkets, when accusations of scamming arise, vendors might modify their language to distance themselves from the behaviours associated with scammers, thereby preserving or rehabilitating their reputations. Vendors accused of scamming may strategically use language that emphasises honesty, reliability, and other trust-building characteristics. They might also avoid certain terms that have become associated with scamming behaviours due to forum discussions or community warnings. This raises the question of how to detect and analyse the absence of language traces. Vendors might also employ counter-allegations or other defensive strategies in their descriptions, directly addressing and refuting scamming accusations, which could change the linguistic patterns observed in their profiles. Such an analysis could improve the understanding of non-violent conflict resolution strategies used by sellers, like negotiation, avoidance, and third-party intervention (Morselli et al. 2017).

6. Conclusion and Prospects for Subsequent Research

In conclusion, the main hypothesis—that it is possible to distinguish reported scammers from legitimate sellers on cryptomarkets using language traces—is refuted by most of the experimental results. This highlights the challenges of using linguistic analysis alone for scam detection in those virtual settings and suggests the need to combine language traces with transactional traces to effectively distinguish between scammers and legitimate vendors. The difficulty of discerning behaviours based on language traces in cryptomarkets can be regarded as a preventive argument aimed at alerting prospective buyers to these platforms. Globally, we observed a pronounced, albeit modest, emphasis on language related to 'Quality and Price', 'Problem Resolution, Communication and Trust',

and 'Shipping'. These findings led us to hypothesise that scammers may frequently provide extensive details about transactions and delivery. This could be a strategic approach to address customers' potential apprehensions, aiming to establish a semblance of trust and reliability in their operations.

Further investigations are, however, needed. A prospective avenue for subsequent inquiry may entail refining the categorisation of vendors identified as scammers in Dread forum posts. Subsequently, future research efforts could explore the possibility of implementing a more nuanced classification schema for these vendors with the intent of distinguishing various typologies of fraudulent behaviours. It would also be interesting to determine whether it is possible to distinguish genuine reviews from fake ones on the Dread forum based on linguistic traces, in order to enrich the research ground.

Our present analytical methodology encompasses the application of machine learning classification algorithms and topic modelling, enhanced by vectorisation techniques like the tf-idf metric. We have experimented with various classifiers, including multinomial Naive Bayes, support vector machines, and Random Forest. Preliminary findings are revealing intriguing aspects, particularly regarding how vendors establish communication channels with buyers. A notable trend is the encouragement of direct contact through encrypted social media platforms such as Telegram and Wickr. These results, while promising, demand a more thorough analysis. The disparity in the volume of documents between reported scammers and other vendors, coupled with the necessity of categorising different types of scammers more precisely, necessitates a cautious approach before drawing definitive conclusions and publication.

With regard to the pre-processing applied to the texts, we made the choice to eliminate stopwords in order to retain only the main words. However, several studies in the literature have suggested that stopwords can be significant elements, and function words have proved useful in previous authorship attribution studies (Arun et al. 2009). It would therefore be interesting to analyse these stopwords in a future study to determine whether they can discriminate between scammers and legitimate sellers.

Moreover, it was observed that a subset of seventeen profiles bore the singular description 'banned'. This unequivocally signifies that the respective vendors associated with these profiles have been banished from the platform. The presence of such data provides some form of ground truth regarding the nature of these vendors. Given the longitudinal nature of the data collection, which spanned an extended timeframe, it is feasible to trace the trajectories of these vendors by examining the evolution of their profiles and listings leading up to their expulsion from the platform. Consequently, we envisage an analysis of the trajectory that could shed light on the developmental patterns of these profiles before the vendors' eventual banishment. Indeed, the adaptive nature of language in vendor descriptions could reflect a complex interplay of reputation management, community interaction, and possibly deceptive strategies. This can be particularly revealing when combined information is extracted from transactional traces, such as the number of sales, the number of won and lost disputes, the score, or the quantities of positive and negative feedback. This information may show a pattern of escalation or changes in behaviour prior to the ban. Future research could benefit from examining these changes over time, potentially applying longitudinal text analysis to capture the evolution of language in response to community feedback and accusations. This would provide a richer understanding of the dynamics at play in cryptomarket ecosystems. It might also be interesting to use a corpus of ads and profiles from legitimate market platforms on the web to see if behaviours, and consequently language traces, differ.

Author Contributions: Conceptualization, C.D., J.L. and Q.R.; Methodology, C.D., J.L. and Q.R.; Validation, J.L. and Q.R.; Formal analysis, C.D. and Q.R.; Investigation, C.D.; Writing—original draft, C.D.; Writing—review & editing, C.D., J.L. and Q.R.; Visualization, C.D. and Q.R.; Supervision, J.L. and Q.R.; Project administration, J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

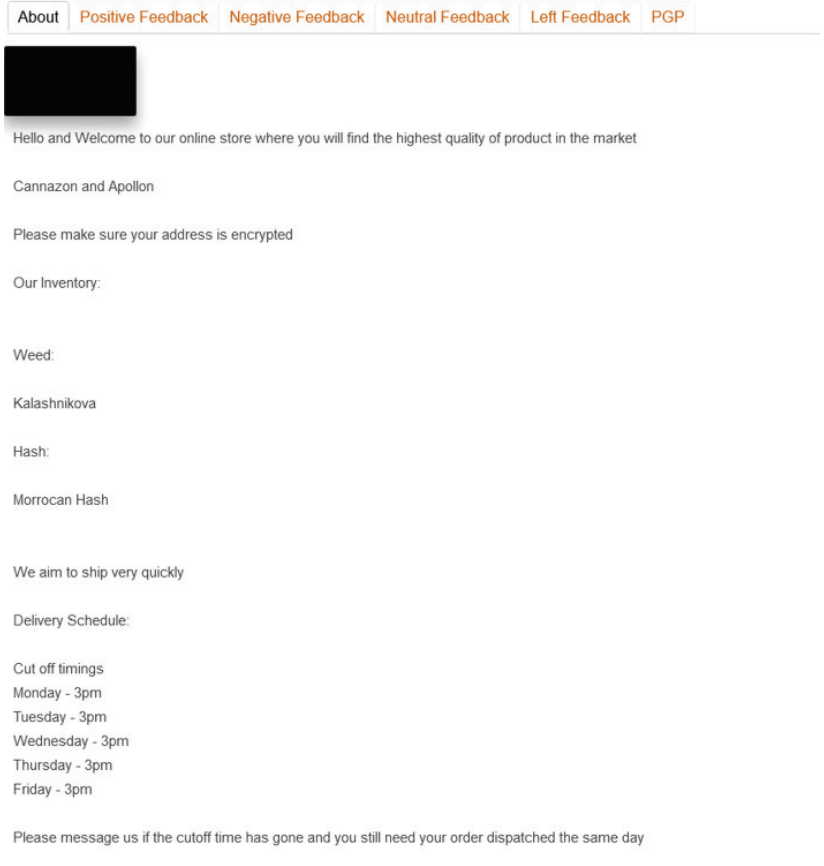


Figure A1. Example of Vendor Profile on Empire Market.


Description Feedback Refund policy

10 Gram * Das Beste Oder Nix*** 92.0% Reines KOKS *****

DEUTSCH

Einfach probieren Leute. Das Zeug ist mega.
Krümmelt wie Kreide. Keine Razierklinge nötig.

DAS BESTE MATERIAL IN DER SCHWEIZ

★PROMOTION★ 10 G KOKS AAA+++ HÖCHSTE QUALITÄT

FLOCKEN, UNBEHANDELT UND 92.0% REIN ★FULL ESCROW★ ★★

★ PROMOTION SALE PRICE

★ 10 Gram
★ Bolivianisches Kokaine
★ Flakes Fishscale
★ REINHEITS GRAD 92.0%
★ Straight from the Brick
★ NIEDRIGSTER PRICE/HÖCHSTE QUALITÄT GARANTIERT
★ 99% of orders arrive in 2 days - Shipping = NUR VON UND IN DIE SCHWEIZ
★ SAUBERES UND ANGENEHMES HIGH
★ STEALTH SHIPPING
★ 5 Euro Shipping and handling A-POST (ABSOLUT NO RESHIP)
★ 10 Euro SHIPPING AND HANDLING Option MIT 100% RESHIP
★★ READ THE REFUND & RESHIP POLICY
★★ FOR FURTHER QUESTIONS CONTACT US!!

ENGLICH

★PROMOTION★ 10 G COCAINE AAA+++ HIGH QUALITY Flakes Fishscale uncut above 92.0% ★FULL ESCROW★ ★★

THE BEST COCAINE IN ALL OF SWITZERLAND

★ PROMOTION SALE PRICE

★ 10 Gram
★ Bolivian Cocaine
★ Flakes Fishscale
★ Purity level above 92.0%
★ Straight from the Brick
★ LOWEST PRICE/HIGHEST QUALITY GUARANTEE
★ 99% of orders arrive in 2 days - Shipping = Only from and to Switzerland
★ Clean and long lasting strong high
★ STEALTH SHIPPING
★ 5 Euro SHIPPING COST A-Post (ABSOLUT NO RESHIP)
★ 10 Euro SHIPPING COST Option WITH 100% RESHIP
★★ READ THE REFUND & RESHIP POLICY
★★ FOR FURTHER QUESTIONS CONTACT US!!

Figure A2. Example of Product Description on Empire Market. Each seller chooses their own layout to highlight the information they wish to communicate to their customers.

Note

[1] https://en.wikipedia.org/wiki/Dread_(forum), accessed 15 November 2022.

References

Addawood, Aseel, Adam Badawy, Kristina Lerman, and Emilio Ferrara. 2019. Linguistic cues to deception: Identifying political trolls on social media. Paper presented at International AAAI Conference on Web and Social Media, Münich, Germany, June 11–14; pp. 15–25.

Aitken, Colin, and Franco Taroni. 2005. Statistics and the Evaluation of Evidence for Forensic Scientists. *Significance* 2: 40–43.

Arun, Rajkumar, Venkatasubramaniyan Suresh, and CE Veni Madhavan. 2009. Stopword graphs and authorship attribution in text corpora. Paper presented at 2009 IEEE International Conference on Semantic Computing, Berkeley, CA, USA, September 14–16; pp. 192–96.

Bancroft, Angus, Tim Squirrell, Andreas Zaunseder, and Rafanell Irene. 2020. Producing Trust Among Illicit Actors: A Techno-Social Approach to an Online Illicit Market. *Sociological Research Online* 25: 456–72.

Bozza, Silvia, Claude-Alain Roten, Antoine Jover, Valentina Cammarota, Lionel Pousaz, and Franco Taroni. 2023. A model-independent redundancy measure for human versus ChatGPT authorship discrimination using a Bayesian probabilistic approach. *Scientific Reports* 13: 19217. [CrossRef] [PubMed]

Button, Mark, Carol McNaughton Nicholls, Jane Kerr, and Rachael Owen. 2014. Online frauds: Learning from victims why they fall for these scams. *Australian & New Zealand Journal of Criminology* 47: 391–408.

Charaudeau, Patrick. 1992. Grammaire du sens et de l'expression. *Hachette*, *epub ahead of print*.

Chaski, Carole E. 2005. Who's At The Keyboard? Authorship Attribution in Digital Evidence Investigations. *International Journal of Digital Evidence* 4: 14.

Christin, Nicolas. 2013. Traveling the silk road: A measurement analysis of a large anonymous online marketplace. Paper presented at 22nd International Conference on World Wide Web, Rio de Janeiro, Brazil, May 13; pp. 213–24. Available online: https://dl.acm.org/doi/10.1145/2488388.2488408 (accessed on 9 November 2023).

Coleman, Meri, and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology* 60: 283–84. [CrossRef]

Degeneve, Clara, Julien Longhi, and Quentin Rossy. 2022. Analysing the digital transformation of the market for fake documents using a computational linguistic approach. *Forensic Science International: Synergy* 5: 100287. [PubMed]

Décary-Hétu, David, Masarah Paquet-Clouston, Martin Bouchard, and Carlo Morselli. 2018. *Patterns in Cannabis Cryptomarkets in Canada in 2018*. Ottawa: Public Safety Canada.

Fobbe, Eilika. 2020. Text-Linguistic Analysis in Forensic Authorship Attribution Forensic Linguistics: New Procedures and Standards. *International Journal of Language & Law* 9: 93–114.

Gibbons, John, and M. Teresa Turell, eds. 2008. *Dimensions of Forensic Linguistics*. AILA applied linguistics series v. 5. Amsterdam and Philadelphia: John Benjamins Pub.

Hancock, Jeffrey T., Jennifer Thom-Santelli, and Thompson Ritchie. 2004. Deception and design: The impact of communication technology on lying behavior. Paper presented at SIGCHI Conference on Human Factors in Computing Systems, Vienna, Austria, April 24–29; pp. 129–34.

Hauch, Valerie, Iris Blandón-Gitlin, Jaume Masip, and Siegfried L. Sporer. 2015. Are computers effective lie detectors? A meta-analysis of linguistic cues to deception. *Personality and Social Psychology Review* 19: 307–42. [CrossRef] [PubMed]

Jacquart, Bérangère, Adrien Schopfer, and Quentin Rossy. 2021. Mules financières: Profils, recrutement et rôles de facilitateur pour les escroqueries aux fausses annonces. *Revue Internationale de Criminologie et de Police Technique et Scientifique* 4/21: 409–26.

Jakupov, Alibek, Julien Mercadal, Besma Zeddini, and Julien Longhi. 2022. Analyzing Deceptive Opinion Spam Patterns: The Topic Modeling Approach. Paper presented at 2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI), Macao, China, October 31–November 2; pp. 1251–61. Available online: https://ieeexplore.ieee.org/document/10097994/ (accessed on 1 May 2023).

Junger, Marianne, Luka Koning, Pieter Hartel, and Bernard Veldkamp. 2023. In their own words: Deception detection by victims and near victims of fraud. *Frontiers in Psychology* 14: 1135369. [CrossRef]

Juola, Patrick. 2021. Verifying authorship for forensic purposes: A computational protocol and its validation. *Forensic Science International* 325: 110824. [CrossRef]

Juola, Patrick, John Sofko, and Patrick Brennan. 2006. A Prototype for Authorship Attribution Studies. *Digital Scholarship in the Humanities* 21: 169–78. [CrossRef]

Longhi, Julien. 2021. Using digital humanities and linguistics to help with terrorism investigations. *Forensic Science International* 318: 110564. [CrossRef] [PubMed]

Markowitz, David M. 2023. Deconstructing Deception: Frequency, Communicator Characteristics, and Linguistic Features of Embeddedness. Available online: https://doi.org/10.31234/osf.io/tm629 (accessed on 9 April 2024).

Markowitz, David M., Jeffrey T. Hancock, Michael T. Woodworth, and Maxwell Ely. 2023. Contextual considerations for deception production and detection in forensic interviews. *Frontiers in Psychology* 14: 1134052. [CrossRef] [PubMed]

Martin, James. 2013. Lost on the Silk Road: Online drug distribution and the 'cryptomarket'. *Criminology & Criminal Justice* 14: 351–67.

Morselli, Carlo, David Décary-Hétu, Masarah Paquet-Clouston, and Judith Aldridge. 2017. Conflict Management in Illicit Drug Cryptomarkets. *International Criminal Justice Review* 27: 237–54. [CrossRef]

Ott, Myle, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 2011. Finding Deceptive Opinion Spam by Any Stretch of the Imagination. Paper presented at 49th Annual Meeting of the Association for Computational Linguistics, Portland, OR, USA, June 19–24; pp. 309–19.

Pavlou, Paul A., and Angelika Dimoka. 2006. The Nature and Role of Feedback Text Comments in Online Marketplaces: Implications for Trust Building, Price Premiums, and Seller Differentiation. *Information Systems Research* 17: 392–414. [CrossRef]

Przepiorka, Wojtek, Lukas Norbutas, and Rense Corten. 2017. Order without Law: Reputation Promotes Cooperation in a Cryptomarket for Illegal Drugs. *European Sociological Review* 33: 752–64. [CrossRef]

Renaut, Laurène, Laura Ascone, and Julien Longhi. 2017. De la trace langagière à l'indice linguistique: Enjeux et précautions d'une linguistique forensique. *Ela. Études de Linguistique Appliquée*, 423–42.

Ribaux, Olivier. 2023. *De la Police Scientifique à la Traçologie*, 2nd ed. Sciences forensiques. Lausanne: EPFL Press. Available online: https://www.epflpress.org/produit/672/9782889155446/de-la-police-scientifique-a-la-tracologie (accessed on 24 October 2023).

Rossy, Quentin, and David Décary-Hétu. 2017. Internet traces and the analysis of online illicit markets. In *The Routledge International Handbook of Forensic Intelligence and Criminology*, 1st ed. Edited by Quentin Rossy, David Décary-Hétu, Olivier Delémont and Massimiliano Mulone. London: Routledge, pp. 249–63. Available online: https://www.taylorfrancis.com/books/9781134888955/chapters/10.4324/9781315541945-21 (accessed on 13 February 2021).

Rossy, Quentin, and Olivier Ribaux. 2020. Orienting the Development of Crime Analysis Processes in Police Organisations Covering the Digital Transformations of Fraud Mechanisms. *European Journal on Criminal Policy and Research, Epub ahead of print*. [CrossRef]

Roux, Claude, Rebecca Bucht, Frank Crispino, Peter De Forest, Chris Lennard, Pierre Margot, Michelle D. Miranda, Niamh NicDaeid, Olivier Ribaux, Alastair Ross, and et al. 2022. The Sydney declaration—Revisiting the essence of forensic science through its fundamental principles. *Forensic Science International* 332: 111182.

Soska, Kyle, and Nicolas Christin. 2015. Measuring the Longitudinal Evolution of the Online Anonymous Marketplace Ecosystem. *24th USENIX Security Symposium*, *Epub ahead of print*.

Taroni, Franco, Paolo Garbolino, Silvia Bozza, and Colin Aitken. 2022. The Bayes' factor: The coherent measure for hypothesis confirmation. *Law, Probability and Risk* 20: 15–36. [CrossRef]

Titus, Richard M., Fred Heinzelmann, and John M. Boyle. 1995. Victimization of persons by fraud. *Crime & Delinquency* 41: 54–72.

Tzanetakis, Meropi, Gerrit Kamphausen, Bernd Werse, and Roger von Laufenberg. 2016. The transparency paradox. Building trust, resolving disputes and optimising logistics on conventional and online drugs markets. *International Journal of Drug Policy* 35: 58–68. [CrossRef] [PubMed]

van Deursen, K. 2021. The Effect of Feedback Polarity on the Sales and Prices on Cryptomarket AlphaBay. Bachelor thesis, Utrecht University, Utrecht, The Netherlands.

Vidros, Sokratis, Constantinos Kolias, Georgios Kambourakis, and Leman Akoglu. 2017. Automatic Detection of Online Recruitment Frauds: Characteristics, Methods, and a Public Dataset. *Future Internet* 9: 6. [CrossRef]

Wright, David. 2014. *Stylistics Versus Statistics: A Corpus Linguistic Approach to Combining Techniques in Forensic Authorship Analysis Using Enron Emails*. Leeds: University of Leeds.

languages

MDPI

Article

In Scriptura Veritas? Exploring Measures for Identifying Increased Cognitive Load in Speaking and Writing

Kajsa Gullberg [1,*], Victoria Johansson [1,2] and Roger Johansson [3]

[1] Centre for Languages and Literature, Lund University, 221 00 Lund, Sweden; victoria.johansson@hkr.se
[2] Faculty of Education, Kristianstad University, 291 88 Kristianstad, Sweden
[3] Department of Psychology, Lund University, 221 00 Lund, Sweden; roger.johansson@psy.lu.se
* Correspondence: kajsa.gullberg@ling.lu.se

Abstract: This study aims to establish a methodological framework for investigating deception in both spoken and written language production. A foundational premise is that the production of deceitful narratives induces a heightened cognitive load that has a discernable influence on linguistic processes during real-time language production. This study includes meticulous analysis of spoken and written data from two participants who told truthful and deceitful narratives. Spoken processes were captured through audio recordings and subsequently transcribed, while written processes were recorded using keystroke logging, resulting in final texts and corresponding linear representations of the writing activity. By grounding our study in a linguistic approach for understanding cognitive load indicators in language production, we demonstrate how linguistic processes, such as text length, pauses, fluency, revisions, repetitions, and reformulations can be used to capture instances of deception in both speaking and writing. Additionally, our findings underscore that markers of cognitive load are likely to be more discernible and more automatically measured in the written modality. This suggests that the collection and examination of writing processes have substantial potential for forensic applications. By highlighting the efficacy of analyzing both spoken and written modalities, this study provides a versatile methodological framework for studying deception during language production, which significantly enriches the existing forensic toolkit.

Keywords: keystroke logging; forensic linguistics; fluency; disfluency; pauses; revisions; planning; language production

Citation: Gullberg, Kajsa, Victoria Johansson, and Roger Johansson. 2024. In Scriptura Veritas? Exploring Measures for Identifying Increased Cognitive Load in Speaking and Writing. *Languages* 9: 85. https://doi.org/10.3390/languages9030085

Academic Editor: Jeanine Treffers-Daller

Received: 26 November 2023
Revised: 7 February 2024
Accepted: 21 February 2024
Published: 29 February 2024

1. Introduction

"To speak the truth is easy and pleasant" was Yeshua's answer when Pontius Pilate asked about his suspected treason towards the Roman Empire in Mikhail Bulgakov's *The Master and Margarita*. Perhaps it is easier to be truthful than it is to lie. However, being able to lie is an important skill to develop in life, for example, to respond politely when your mother-in-law asks you if you like her horrendous stew (Talwar 2019). Despite the social function of some lies, there are also instances when it is important to be able to tell if a person is lying or not, in our day-to-day lives, as well as in our legal system, where, for example, witness accounts need to be judged for their credibility in a safe and just way. While there have been attempts at creating a "lie detector", so far these efforts have not reached a reliable and safe conclusion and no single "symptom" of a lie has been identified that can be used for diagnosing a story as deception (Ofen et al. 2017; Mann 2019; Vrij et al. 2022).

Instead, discriminating a lie from the truth is typically contingent upon a comprehensive evaluation of multiple distinct verbal and nonverbal behavioral indicators, such as gaze cues, pulse rate, hand movements, and manifestations of nervousness, among others (Newman et al. 2003; DePaulo et al. 2003; Vrij et al. 2010; Granhag et al. 2015). The present methodological article sets out to contribute to the existing body of behavioral indicators.

We suggest that the examination of linguistic processes that occur during spoken and written language production, captured through audio and video recordings for speaking and keystroke logging for writing, may offer new possibilities in forensic linguistics. Recordings of spoken reports have previously been used in deception studies, for example, to identify verbal cues of deception (Vrij et al. 2022), and these often combine production cues such as hesitations with content cues such as the linguistic complexity of a message. Complementing these verbal cues with studies and methods examining both spoken and written real-time production will further the understanding of deception, especially with the use of the knowledge that can be gained from the existing body of linguistic research on language production in both speaking and writing.

As mentioned above, cues in speech are already one aspect examined in deception studies and detection, and deception during writing has also been studied (e.g., see Banerjee et al. 2014; Derrick et al. 2013). These studies have, however, not used the full potential and tools provided by the last decades' research in the area of cognitive writing processes (Lindgren and Sullivan 2019), which has developed methodologies and informed theories on how the writing process unfolds in real-time through the study of pauses and revisions, for example.

Knowledge about writing in itself is essential in modern society, where writing is a common and often necessary form of communication in a range of activities for a majority of the population (Brandt 2015). As such, while many reports in forensic settings are spoken, written reports are also becoming more and more common. For example, in a society marked by an escalated reliance on digital platforms for diverse purposes, individuals are increasingly prompted to report accidents, incidents, and criminal activities online. These reports find their outlets on platforms such as the websites of national law enforcement agencies or insurance providers. Furthermore, written narratives are employed by migration authorities and in the creation of ongoing military reports. Consequently, it becomes conceivable to use software designed for capturing real-time text production, similar to how spoken processes are recorded through audio and video. Recent advancements in web-based software now afford such capabilities. Several prototypes have been developed, with a notable example being the keystroke logging tool, CyWrite, which resembles a customized web-based text editor that has the capacity to comprehensively capture and later reconstruct the entire text production process (Chukharev-Hudilainen et al. 2019). The prospect of scrutinizing real-time writing processes for forensic purposes, paralleling the examination of real-time spoken processes, holds significant intrigue. It is important to note, however, that this application is unlikely to function as prima facie proof of lying, or as a reliable "lie detection". Nevertheless, it cautiously holds the potential to identify circumstances warranting further attention during subsequent interrogations.

When addressing lying and deception in language production, it is imperative to acknowledge that behavioral indicators may be attributable to factors other than the veracity of the communicated information. For instance, it is easy to imagine that in an interrogation situation, most witnesses of a potential crime will feel an obligation to provide accurate information and that the gravity of the situation will bring on nervousness. The challenge lies in discerning such reactions from those exhibited by individuals who are anxious about the detection of a lie.

But what is a lie then? Lies can take on many different forms and can even be defined differently across languages and cultures—like whether or not the *intention to lie* or *the factual truth* is the defining feature of a lie (Coleman and Kay 1981; Nishimura 2018). A person can lie by omission, i.e., by leaving out important information, or they can lie by presupposition ("Has Bambi stopped hitting Thumper?" presupposing that Bambi has been hitting Thumper). In addition, different kinds of misleading information can be considered lies. In the scope of this article, lying is operationalized as a person *knowingly* giving false information with the *intention* of making the receiver of this information believe it to be true. Deception may manifest in both spoken and written modalities, impacting various *linguistic processes*, such as *planning, conceptualizing, generating spoken or written text,*

monitoring, and *editing,* irrespective of the communicative mode employed (Flower and Hayes 1981; Levelt 1989).

The present study forms part of a larger project entitled "Based on a true story: How to differentiate between invented and self-experienced narratives through comparing linguistic processes in speaking and writing.". The project is guided by the foundational assumption that heightened cognitive load during the process of language production, whether in spoken or written form, is likely to exert a discernible influence on linguistic processes (Goldman Eisler 1968; Matsuhashi 1981; McCutchen 1996). The overarching objective of this larger project is to examine how (and if) deception influences cognitive load during language production and to ascertain whether discerning deception is more realizable through an examination of writing processes as opposed to spoken processes. To achieve this goal, the initial step involves identifying and defining relevant phenomena that can effectively capture and measure heightened cognitive load during language production across these two different modalities.

The present study aims to accomplish this first step by establishing a versatile methodological framework, capable of identifying and quantifying phenomena linked to heightened cognitive load in both spoken and written language production. We believe such an approach holds the potential to significantly expand the existing forensic toolkit.

1.1. Deception and Cognitive Load

The relationship between deception and increased cognitive load has been the subject of extensive investigation in various studies. Cognitive load refers to the demand placed on working memory resources when solving immediate tasks (Baddeley 2007; Cowan 2010). The underlying assumption is that lying is a mentally demanding task, prompting suggestions that indicators of heightened cognitive load could be employed for deception detection. All forms of lie detection rely on individuals perceiving deception and identifying cues (either automatically or manually) that may suggest falsehood. A myriad of lie detection techniques has been proposed, particularly for use in interrogations and interviews (see Walczyk et al. 2013 for a comprehensive overview). These techniques may focus on attentional processes, aligning with the orienting response theory (Sokolov 1963), or delve into memory processes and inhibition, in line with the parallel task set model (Seymour 2001).

Several theoretical frameworks address deception and its relation to cognitive load. One is the four-factor theory of deception, advanced by Zuckerman et al. (1981), which posits that deception escalates cognitive load. Some theories include explanations as to why deception would increase cognitive load. For instance, the interpersonal deception theory (Buller and Burgoon 1996; Burgoon and Buller 2008) proposes that cues of deception stem from aspects of communication that remain "unmonitored" due to increased cognitive load. Another example is the self-presentation theory (DePaulo 1992), which outlines three cognitive phases governing behavior to appear truthful: intention to regulate behavior, intention translated into non-verbal behavior, and self-assessment of the behavior. Sporer and Schwandt (2006, 2007) introduced the Working Memory Model of Deception, which builds on Baddeley's (2007) working memory model and asserts that lying elevates cognitive load, potentially affecting speech production among other processes. Finally, the Activation-Decision-Construction Model (Walczyk et al. 2003, 2005, 2009) outlines a model for deceptive responses in the context of lie detection interviews, and this model has been expanded to account for repeated lies (2009).

These theories and models have undergone scrutiny, as exemplified by a study conducted by Repke et al. (2018) that tested two models—one assuming that increased cognitive load during deception would reduce linguistic complexity and another assuming that the lie's goal would determine the complexity of the deception. The latter model received empirical support, indicating that liars can adjust the complexity of their falsehoods based on their objectives. Other studies propose content analysis to assess statement credibility, such as criteria-based content analysis (CBCA) (Vrij et al. 2000), developed to evaluate state-

ments from individuals who have experienced abuse, as well as analyses of vividness and spontaneity in lies versus truths (Colwell et al. 2007). These investigations have revealed that the content of statements is, to some extent, influenced by whether a person is lying or telling the truth, and complementing this knowledge of the content with future knowledge about the process of deceiving would most likely be rewarding.

Similarly, Leins et al. (2012) explored the impact of different reporting modes on deception from a forensic perspective. They discovered that, when individuals were asked to recount the same event through spoken and pictorial modes, liars exhibited less consistency across the two modes compared to truth-tellers. Thus, while liars can tailor their lies to specific goals in a given context, transferring lies across different reporting modes appears more challenging. Further investigation into this phenomenon across various language modalities could be valuable. Additionally, Vrij et al. (2008) found that, when asked to narrate a series of events in reverse order, both verbal and nonverbal cues indicative of deception (e.g., filled pauses, hesitations, and leg movements) increased among liars. Moreover, they observed an improvement in lie detection accuracy among police officers in the reversed condition, surpassing chance levels.

Regarding heightened cognitive load during deception, reaction time studies have been a common approach. For instance, Duran et al. (2010) reported a significant increase in reaction time when participants were instructed to lie in response to simple yes/no questions, a pattern consistent with numerous other studies (e.g., see Suchotzki et al. 2017 for a review of studies measuring reaction time in relation to deception). Furthermore, Debey et al. (2012) found that, when given additional time to examine a stimulus after being instructed to lie, participants exhibited significantly longer reaction times when lying compared to responding truthfully.

Many studies investigating deception cues primarily focus on interview responses or spoken language. Apart from reaction time latency, some also assess speech rate and hesitations as verbal cues of deception. For instance, Vrij et al. (2008) examined speech rate (calculated as the number of words divided by the length of the answer) and found that liars had a slower speech rate than truth-tellers, along with more hesitations. A limited number of studies have specifically looked at deception related to aspects of speaking and writing. One example is the study by Goupil et al. (2021), who found that the speech signal itself may be perceived as more or less honest. A few recent examples include studying deception during written language production, with interesting findings, such as liars engaging in more revisions and producing shorter texts (Banerjee et al. 2014). Another result has demonstrated that, in synchronous chat settings, liars exhibit not only increased revisions and shorter texts but also longer response times (Derrick et al. 2013). The latter study also noted a significant age-related effect, with older participants displaying these behaviors more prominently than younger individuals. Finally, studies have demonstrated that writing processes can be disrupted due to background speech, especially regarding semantic aspects, something that may be relevant for inducing increased cognitive load in experimental settings (Sörqvist et al. 2012).

In sum, the connection between cognitive load and deception is underpinned by theoretical models as well as empirical findings. The review of the field further highlights that the examination of behavior during the production of language may add insights into when and how deception occurs.

1.2. Language Production and Cognitive Load

The concept of working memory has also been influential in descriptions of language production. Across all kinds of language production, we depend on our working memory resources to perform tasks, such as planning what to say/write, actually expressing it, and evaluating the result of it (McCutchen 2000; Baddeley 2007). In studies on speaking and writing, working memory demands, or *cognitive load*, have been investigated through analyses of pauses and disfluencies (cf. e.g., Goldman Eisler 1968; Matsuhashi 1981; Spelman Miller 2006b). The underlying idea posits that, when too much information needs

to be processed simultaneously, our limited working memory capacity becomes overloaded with information and additional time is needed to plan for spoken and written expressions. This often results in longer pauses and/or more frequent pausing (Spelman Miller 2006a), alongside an increased occurrence of other expressions of disfluencies in speaking, such as filled pauses, elongated words and word segments, and repeated words and expressions (Goldman Eisler 1968; Clark and Wasow 1998; Heldner and Edlund 2010), and in writing, such as deletion of word fragments, words, and expressions, additions, and substitutions, both locally and globally.

Numerous factors may contribute to cognitive load during language production (for examples and overviews addressing different factors influencing cognitive load see Barkaoui 2019; Bourdin and Fayol 1994; Feng and Guo 2022; Johansson 2009; Kellogg 2008; Kellogg et al. 2016; Lively et al. 1993; Lourdes Ortega 2009; Manchón 2020; Song and Li 2020). Existing research in this area suggests that writers' and speakers' linguistic proficiency (including factors such as producing in one's first or second/third language, as well as overall grammatical and lexical knowledge), age, and education will influence fluency during language production. In addition, factors such as knowledge of genre, topic, and the amount of preparation, as well as grammatical complexity will have an impact. Finally, contextual factors such as sleep, hunger, general comfort, distracting factors in the current situation, etc., will sway the performance. In sum, differences in execution at the group level are expected; for example, first-language speakers are generally more fluent than second-language speakers, or increased age and education lead to more fluent speaking and writing (compared to children). Apart from this, overall between-subject findings and substantial within-subject findings can also be expected. That is, a person does not always pause, revise (in writing), or demonstrate disfluencies (in speaking) in a consistent way. The context will matter in this respect. For all these reasons, it is necessary to establish an individual baseline or include a control condition in all research using the assumed symptoms of cognitive load during language production (such as pausing and revision) as indications of—in this case—lying.

Thus, with all things being equal, in this case, it is assumed that, when a person engages in deceptive communication, an augmented cognitive load can be reflected in the language production processes. As mentioned, reaction time studies on deception have shown that both prompted and unprompted lies lead to an increase in response time during lies (Duran et al. 2010; Debey et al. 2012; Williams et al. 2013; Walczyk et al. 2013; Suchotzki et al. 2017; Bott and Williams 2019). Increased cognitive load during lying would be caused by the speakers/writers having to concurrently devise *what* to say/write and determine *how* to say/write it: organize the sequence of (in this case) narrative events, select an appropriate syntactic structure, and choose the lexical items. Additionally, the speakers/writers would have to actively maintain a mental representation of the truthful version of the events and continuously decide *when* and *how* they wish to deviate from this. All these factors add to working memory demands.

1.3. Speaking and Writing

Language production is an over-arching term for the modalities of speaking, writing, and using sign language (which will not be addressed here). Below, we outline some fundamental characteristics that apply to spoken and written discourses and that will influence the behavior of speakers/writers. Here, we disregard situations such as instant messaging or fast-written conversations, spoken conversations on the phone, or messages in delayed mode (i.e., voice mail and voice messages), where some of the characteristics of the modality will be less prominent. We use a contrastive focus and address the three themes: *time*, *receiver*, and *permanence*.

1.3.1. Time

The difference in production speed is an essential factor for understanding how speakers and writers distribute their resources during language production. The rate at

which language can be produced in the two modalities is profoundly different; we speak much faster than we write. A common estimation is that speaking (in English) allows for a speed of 120 to 200 words per minute, corresponding to approximately 2–10 syllables or 8–15 phonemes per second (Crystal and House 1990; Schreiber and McMurray 2019), while a proficient typist would produce 38–40 words per minute (Hayes and Chenoweth 2006). The differences can mostly be attributed to the fact that speaking "only" requires the use of the vocal apparatus for expression, whereas writing requires the use of some artifact, e.g., pen and paper, keyboard, screen, etc. The writers' mastering of the artifact and often, the limitation of the artifact itself, introduce an intrinsic latency in the transformation from one's thoughts into linguistic expression (Grabowski 2008).

Finally, in the context of speaking, there are one or more listeners waiting for the delivery of the message and they can potentially interrupt. As a result, speakers will often experience time constraints and consequently use strategies that allow them to plan what to say while keeping the floor. This encompasses the use of filled pauses (*eh, um*) to indicate that more will come, to repeat and reformulate words and phrases, and to allocate silent pauses within syntactic units. These strategies are typically learned very early in life through numerous interactions and observations of spoken contexts, and speakers are very rarely aware of this behavior. The context for writing is normally different: even in stressful situations, writers will have comparably more time to think, generate text, and edit it before handing it over to a reader.

1.3.2. Receiver

Another substantial distinction between the modalities is the presence or absence of a receiver, or in other words, a listener or reader (Chafe and Danielwicz 1987; Chafe and Tannen 1987). Speakers often rely on the listeners' reactions to determine if, and when, more information is required (Levelt 1989; Barker et al. 2020). Conversely, writers must anticipate the readers' knowledge and needs and tailor the text accordingly. This inherent uncertainty can result in extensive revisions during and after text production (Flower and Hayes 1981; Hayes et al. 1987). However, these revisions and alterations are usually not visible to future readers, something that can be contrasted to the spoken context where listeners will be aware of all modifications made during speaking. Speaking is thus described as a dialogic activity, while writing is characterized as monologic (Linell 2009).

1.3.3. Permanence

The visibility and (relative) permanency of the written message in relation to the fugitive nature of the spoken message is yet another factor to consider. The fleeting spoken discourse necessitates repetition if the speakers need to reinforce certain points (Levelt 1989; Clark 1996). The listeners can also readily discern hesitations and repetitions, phenomena that help signal that the turn is ongoing and that the speakers want to keep the floor (Norrby 2014). Importantly, studies suggest that these disfluencies also facilitate the understanding of the spoken message (Clark 1996; Fox Tree 2001; Fox Tree and Schrock 2002). Conversely, when readers encounter a written text, it typically lacks visible traces of prior revisions.

The permanent condition of the written message is one contributing factor to the higher status of written language (cf. Chafe 1994). The visible language and the delay between the written production and the readers' reception also lead to a strong cultural expectation (reinforced by, e.g., the importance of writing skills posed by formal education and schooling) that the message should be edited and perfected before being handed over (cf. the different functions of spoken and written language outlined by, e.g., Biber 1988; Halliday 1985). The permanency is probably also a fundament for our view that written agreements and contracts are more reliable and binding than (undocumented) oral equivalents: it is more difficult to prove what was said than what was written. Writing thus includes an important component of understanding that one can, and is expected to, revise the written text before it is finished and that the revisions will be (largely) invisible to

readers, especially for computer-written texts (Einarsson 1978). This belief would contribute to writers' revision behavior and how it is distributed throughout a writing session.

In summary, on the one hand, speaking is characterized by its quick, instantaneous, and synchronous nature, with (typically) present listeners who witness the entire overt language process and can actively contribute to the spoken message through oral and visual feedback (e.g., nodding or asking questions). On the other hand, writing typically unfolds more slowly, in isolation without readers present. The written message needs to be decontextualized and is read later and (often) in a different place, which, in turn, requires that writers anticipate what the future readers need to understand the context. The characteristics of spoken and written discourse will influence the type of processes that can be observed during language production in the two modalities.

1.4. Models of Language Production in Speaking and Writing

The description of spoken and written modalities outlined above have been addressed in theoretical models attempting to identify and sequence the different processes involved in speaking and writing. Overarching models, covering both speaking and writing are hard to find (but see Cleland and Pickering 2006). Instead, we use some of the most seminal models for speaking and writing to establish a terminology for the most described processes during language production (note that the description is delimited to stage models):

Models of speech production (e.g., Fromkin 1973; Levelt 1989) distinguish between different stages of production that unfold (somewhat) successively. In a simplified description of the process, it commences with *conceptualization*, where the content is decided, followed by *sentence formation*, where lexical decisions are made, and *syntactic structuring*, determining word order, and ultimately *articulation*. During articulation, the speaker is also engaged in constant *monitoring* of what they are saying, as well as being attuned to the listener's reaction (while at the same time moving on with the production of the next utterance).

Other theories of spoken production emphasize that the speech process is facilitated by certain mechanisms. For instance, Linell (2009) highlights that grammatical constructions used in speech have been internalized by the speaker through prior practice in various situations (cf. Clark 1996 who described how much is given in a conversation, for example, in question-answer constructions). Despite these facilitating mechanisms during speech production, the task of having to *plan what to say while saying it* can still be daunting.

Models of written production (e.g., Flower and Hayes 1981) distinguish between three main processes: *planning*, *translating*, and *revision*. Planning entails the formulation of ideas, text organization, and text generation (on a conceptual and linguistic level), while translating involves rendering these ideas into their orthographic form. Revision encompasses reading and evaluating the text to align with the writers' intended goals, and based on this, editing the text if necessary. These processes are iterative and recursive during the unfolding of text production; thus, one should understand the processes of planning, translation, and revision to be carried out at a local level within short time frames, although they could also be applied for understanding the writing process at a more global text level and on a long-term perspective.

Revisions in writing can take place at any point during the writing process, whenever the writers see fit. This flexibility means that revisions can occur at various locations within the text; writers may edit at the leading edge of the text they have produced so far or make changes at some other point in the text they have already written (Lindgren et al. 2019).

In sum, the theoretical frameworks for speaking and writing, despite differences in terminology, include some common components, such as those of *pre-activities*: conceptualization, sentence formation and syntactic structuring in speech, and planning in writing, those of *text generation*: articulation and translating, and finally, those of *post-activities*: monitoring and revision. Of these processes, it is only the articulation and translation components that are overt and observable, while the other processes need to be inferred or rely on self-reporting methods.

1.5. The Present Study

The review of previous research demonstrates that the examination of symptoms of cognitive load in different behaviors can be fruitful for identifying deceit and for forensic purposes. However, one conclusion is that there is no single action or expression that has been shown to be indicative of lying. Instead, co-occurrences of phenomena may be a way forward. For this reason, it is imperative to increase the forensic toolkit and expand the possibilities to evaluate witness statements and narrative reports. Given the research outlined above, the current study builds on, on the one hand, studies of deception and the effect of cognitive load, and on the other hand, the existing knowledge on language production processes and how demonstrations of increased cognitive load are expressed.

As mentioned in the introduction, the present study has a methodological aim, closely related to a larger research project, called "Based on a true story: How to differentiate between invented and self-experienced narratives through comparing linguistic processes in speaking and writing." The overarching objective of the project is to examine the impact of deception on cognitive load during language production and to investigate if this heightened cognitive load is more salient in writing than in speaking. The first step in reaching this goal is, as mentioned in the introduction, to identify relevant phenomena during language production that can effectively capture and measure heightened cognitive load across these modalities. Thus, the aim of the present study is to accomplish this initial step by establishing a versatile methodological framework capable of identifying and quantifying phenomena associated with heightened cognitive load in both spoken and written language production.

Four central assumptions underpin our approach to achieving this objective. First, acts of deception are mentally demanding, leading to increased cognitive load, which directly affects sensory and behavioral expressions (cf. Vrij et al. 2000; Walczyk et al. 2013), such as prolonged response latencies (e.g., Duran et al. 2010). Second, heightened cognitive load during language production, regardless of modality, will directly impact linguistic processes during real-time language production (Goldman Eisler 1968; Matsuhashi 1981; McCutchen 1996). Third, acts of deception will have a discernible influence on these linguistic processes (Banerjee et al. 2014; Derrick et al. 2013; Vrij et al. 2008). Fourth, language users are generally more familiar with real-time spoken interactions compared to the real-time process of writing. As a result, mimicking truthful spoken language production is likely easier than replicating the same in the context of writing processes.

The complete corpus of the larger project comprises experimental data where, in total, 40 participants recounted events depicted in four specially tailored films portraying minor misdemeanors. In each data collection session, every participant produced four narratives based on four different films: one truthful spoken account, one deceitful spoken account, one truthful written account, and one deceitful written account. Participants returned three times for additional sessions, two weeks apart, and repeated all their narratives. Thus, the complete corpus consists of both truthful and deceitful narratives, totaling 320 written and 320 spoken accounts, collected from the 40 participants. In the truthful condition, participants were asked to retell the events as they unfolded, while in the deceitful condition, participants were asked to modify "who did it". To eliminate potential order effects arising from modality (writing > speaking or speaking > writing), or veracity (truthful > deceitful or deceitful > truthful) variations, a Latin Square design was implemented during data collection. Consequently, all possible orders were equally distributed and counterbalanced across participants throughout the entire data corpus. This design allows for within-subject comparisons across tasks while controlling for potential idiosyncratic veracity and modality effects, but also for the examination of possibly generalizable patterns through between-subject comparisons. The study design further enables an exploration of the dynamics of cognitive load indicators throughout a narrative account and investigates whether specific sequences, particularly those associated with altered portions of the accounts, exhibit distinct patterns. The data within the larger project are extensive, necessitating a systematic approach to its exploration for the present study's purposes.

The present study uses a subset of data samples derived from the corpus collected within the larger project to qualitatively illustrate how language production unfolds in speaking and writing. The critical goal of this description is to identify phenomena that effectively discern and quantify heightened cognitive load—something that will inform the exploration of the corpus in the larger project.

To summarize, in the present study, we use a subset of data samples to qualitatively depict the unfolding of language production in both spoken and written contexts, with the aim to identify observable phenomena that can effectively discern and quantify heightened cognitive load. It is important to note that our study is not an attempt to comprehensively demonstrate how deception affects cognitive load and the associated linguistic processes within spoken and written narratives. Instead, our primary focus is on the development of a methodological framework that would enable such examinations. To achieve this objective, we employ overarching theories related to cognitive load and deception, grounding our analysis in a linguistic approach to interpret real-time language production in both speaking and writing.

Specifically, this study is guided by the following research questions:

1. How can text length be measured to capture increased cognitive load during language production in speaking and writing, respectively?
2. How can pauses be defined and measured to reflect increased cognitive load during language production in speaking and writing, respectively?
3. How can changes in spoken and written text, such as repetitions and reformulations in speaking, and revisions in writing, be defined and measured to reflect increased cognitive load in speaking and writing, respectively?
4. In the assessment of spoken and written language production, how can we use these measures to capture fluctuation in cognitive load throughout the process?

2. Materials and Methods

This section describes the data that are used to exemplify the methodological discussion that constitutes the main part of this study. Note that, given the methodological scope of this study, the measures that are described and discussed in detail in the Results section are only briefly explained and addressed here.

2.1. Materials

The data used in this study were drawn from two participants from the larger project. These participants were randomly chosen with the purpose of providing examples that qualitatively illustrate the methodological aspects of analyzing and measuring behavioral phenomena in language production, which are outlined in the Results section below. In choosing the texts that provide these empirical examples, we selected two participants who had truthfully or deceitfully described the same films: one participant for spoken accounts and one for written accounts. This selection was made in order to restrain the number of film events that are related in the examples, and in that way, hopefully, facilitate the actions and choices of the participants.

The four selected texts from the two participants are illustrative of all the texts in the larger corpus. Note that the excerpts are used to exemplify the behavioral phenomena, and in that sense, they constitute empirical evidence. However, the four texts are not used for any conclusions on a more general level, neither for what is more common during speaking or writing nor for what characterizes deceitful or truthful accounts.

The texts from the two participants comprise spoken and written samples obtained during an experiment in which participants described events they had observed in four elicitation films depicting minor misdemeanors (e.g., cheating on an exam and putting pepper in a stranger's coffee mug). The data were collected online via the Zoom platform due to the constraints imposed by the COVID-19 pandemic, and the participants consequently made all recordings on their own computers and then transferred files to the researcher through a safe server depository. The two participants in the present study

will be referred to using the pseudonyms Alfa and Bravo. Alfa and Bravo were native Swedish speakers with no known difficulties in reading, writing, or speaking, retelling the narratives in Swedish. They had accomplished a minimum of one year of academic studies at the university level. They further demonstrated fairly good typing skills; thus, their transcription skills were not expected to intimidate other engagements during text production (cf. Van Waes et al. 2021). All texts were written on a computer.

The sample included one deceitful and one truthful account from each participant. Alfa's truthful spoken account and Bravo's truthful written account describe the same film (*The Garden Café*), and likewise, Alfa's deceitful spoken account and Bravo's deceitful written account relate to another film (*The Examination*). In the deceitful accounts, Alfa and Bravo were instructed to alter the attribution of the culprit in the videos to simulate a fabricated eyewitness account. In effect, this task entails altering a specific portion of the events in the film, rather than the entirety of the film, although the participants were free to make changes where they saw fit. See the Supplementary Materials for the full transcripts of the spoken data, as well as the final and linear texts of the written data.

Thus, all four narrative text examples describe the events in one of two films, tailored for this experiment. These wordless films were all approximately three minutes long, and each film included three protagonists and depicted a misdemeanor where one protagonist was the culprit. In the deceitful condition, there was always an option of putting the blame on another protagonist. A short synopsis of these two films is given below. The first film, *The Garden Café*, starts with a girl sitting at a table engrossed in her reading, her bag placed on a chair beside her. In the foreground, there is a table with cakes and drinks as well as payment instructions so that guests can serve themselves. A second girl enters and looks around to see what the café has to offer. Shortly thereafter, a man arrives and, without apology, stumbles on the chair with the bag, and the bag falls from the chair, with its contents spilling out. The man then walks up to the tables and checks out what is offered at the café. He cuts in before the other girl has a chance to take what she wants and starts serving himself without paying. When his coffee is left unsupervised, the girl whose bag he knocked over, takes the opportunity to add pepper to it with a pepper mill. She then takes her coffee and leaves the café, and soon after, the girl who he cut in front of also leaves. The man drinks the coffee and reacts to the strange taste by spitting it out. He is unaware of who or what caused this unexpected taste.

The second film, *The Examination* shows a situation where a student cheats during an exam at the university. Two girls are sitting on opposite sides of an aisle in a classroom, and a male teacher is supervising them at the front. We see the backs of the girls: one in a yellow sweater on the left and one in a black sweater on the right. Seizing an opportunity when the teacher's coffee cup falls over and he is busy wiping it up, the girl in the black sweater pulls out a cheat note and looks at it. The girl in the yellow sweater notices what she is doing and observes her. When the exam is over, both girls hand in their exams, and when doing this, the girl with the black sweater accidentally drops her cheat note without noticing. The teacher subsequently finds it on the floor but remains unaware of its owner.

2.2. The Written Data: Data Collection Methods

The written data were collected by means of keystroke logging, using the software ScriptLog, version 196. Keystroke logging is a methodology that enables the study of real-time writing processes, and the software typically collects written data through the registration of keypresses and mouse movements during text production (Wengelin and Johansson 2023). ScriptLog records the writing process and then allows researchers to replay the writing session and extract information regarding where and how long writers pause, what, when, and how much was deleted from the text, and overall statistics of writing time and text length as the number of characters (letters, numbers, punctuation, and spaces) during a writing session. In addition, the software produces various output files, for instance, a *final text*, that is, a version of the finished written text in the way the

writers intended it for the reader and—as a contrast—a so-called *linear text* that illustrates the writers' step-for-step writing process, with pauses and revisions included.

Deciding what constitutes a pause during computer writing has proven to be difficult. Generally, pauses have been defined as "inactivity between keypresses" (Van Waes and Leijten 2015). This inactivity can be measured in milliseconds, meaning that it is detailed enough to measure the writer's transcription skills. However, pause measures are also restricted to the hardware's (computer and keyboard) ability to capture and register keypresses (Johansson et al. 2023). Researchers interested in high-level cognitive processing (such as planning, reading, and revision) during writing have often made use of *pause thresholds*, which filter the noise of low-level transcription processes (such as the transcription speed of the writer or processes such as remembering how to spell certain words) (Wengelin 2006). While an objectively defined pause threshold has yet to be made (see Barkaoui 2019), various ad hoc criteria have been proposed. For instance, 2 s has been established as an ad hoc threshold that will capture most high-level processes while disregarding pauses caused by motor activities (at least for adult writers who are good typists and lack reading and writing difficulties) (Wengelin 2006). However, in studies where smaller fluctuations in pause behavior may be important, it would be safer to postulate a lower threshold to avoid filtering out too many details of the writing process. For this reason, we have chosen the ad hoc criteria of the 1 s threshold. Finally, it should be noted that various pause criteria can easily be explored using the analysis options in keystroke logging programs such as ScriptLog and Inputlog (Leijten and Van Waes 2013).

A comparison of these two outputs from the final and linear texts illustrates the various operations (and the amount of time and effort) writers engage in during the writing of a particular part of the text. Table 1 includes examples of the final text (in the upper part of the table) and its corresponding linear text (in the lower part). In the linear text, actions such as using the 'backspace key' are indicated within angular brackets, as are the occurrence of pauses. In this example, we have included pauses longer than 1 s. Note that, although the linear file provides rich information, it is not always particularly fitted for understanding exactly what has been deleted or comprehending the writers' movements within the text. For such purposes other types of output options can be acquired from keystroke logging software, for instance, revision analysis and possibilities to replay the writing session (see examples from Inputlog, Leijten and Van Waes 2013).

In Table 1, we contrast the final text with the linear text from Bravo's written truthful account to illustrate the type of data that the linear text can provide. As shown in the table, the final text starts by saying "När tjejen hade plockat ihop alla sina saker gick hon fram till kaffebordet och tog upp en pepparkvarn" (*When the girl had collected all her things, she walked up to the coffee table and grabbed a pepper mill*). From the linear text, however, it becomes evident that Bravo started out writing this part of the text differently. At the start of the linear text, there is a pause of 3.250 ms. This is followed by the writing of "När kille" (*when the gu*), that is, starting by telling the events from the perspective of 'the guy'. But then, Bravo presses the backspace 10 consecutive times to delete what has just been written. After that she writes the letter "J", and then immediately deletes that. Instead, she writes "Tjejen" (*The girl*), which is also deleted at once through seven presses on backspace. After this operation, Bravo writes "När tjejen hade plockat" (*When the girl had collected*) which corresponds to the solution in the final text.

This contrasting example of the final and linear texts shows that writers often engage in many more activities during text production than what is visible or traceable from inspecting only the final text. Through the study of such actions, it is possible to explore how writers allocate their resources during writing and gain general insights into where writers need to pause or revise and more specific awareness if this need fluctuates in regard to certain contexts, or in our case, particular sequences. In this study, we are interested in methods for exploring the linear files. Analyses of the final texts are not discussed here, but they can be investigated through corpus linguistic methods or discourse analysis. Future methodological avenues can further encompass comparisons between the linguistic

properties of the final texts and the writing processes, as shown in linear files, which took place during the production of the final texts.

In the Results section, we will refer to translated excerpts in our examples in this study (for full versions of the linear and final texts, see the Supplementary Materials).

Table 1. Written data Example of a written final text and linear text of a truthful account. The excerpt comes from the central segment of a text by Bravo and is a description of an event depicted in one of the elicitation films. The top section of the table shows the *final text*, such as it was when Bravo finished writing. The *linear text* is presented below, which shows the same part of the text but with all pauses and revisions, denoted within angular brackets. The English translation mirrors the syntax and structure of the original Swedish, and keeps any mistakes in the texts.

Final text of written truthful account from Bravo	**English translation**
När tjejen hade plockat ihop alla sina saker gick hon fram till kaffebordet och tog upp en pepparkvarn och hällde peppar i mannens kaffe. Efter det gick ut därifrån. Den yngre tjejen såg detta och gick snabbt iväg från café.	When the girl had picked up all her things she walked up to the coffee table and grabbed a pepper grinder and grinded pepper in the man's coffee. After that walked out from there. The younger girl saw this and quickly left the café.
Linear text of written truthful account from Bravo	**English translation**
<3.250>När kille<BACKSPACE10>J<BACKSPACE1> Tjejen <1.471><BACKSPACE7>När tjejen hade plockat ihp<BACKSPACE1>op alla sina saker gick hon fram till kaffebordet och <1.287>tog upp en pepparkvarn som hon <3.799>använde<BACKSPACE15>och <1.993>hällde peppar i mannens kaffe. Den lilla tj<BACKSPACE8>yngre tjejen <1.372>såg detta och gick <4.163>snabbt igäv<BACKSPACE3>väg från k<BACKSPACE1>café, <BACKSPACE2> vilket <2.792><BACKSPACE8>. <1.505><MOUSECLICK><2.931>Efter det gick ut därifrån och <BACKSPACE5>.	<3.250>When gu<BACKSPACE10>J<BACKSPACE1> The girl <1.471><BACKSPACE7>When the girl had picked pu<BACKSPACE1>up all her things she walked up to the coffee table and <1.287>grabbed a pepper grinder that she <3.799>used<BACKSPACE15>and <1.993>grinded pepper in the man's coffee. The little gi<BACKSPACE8>younger girl <1.372>saw this and walked <4.163>quickly agay<BACKSPACE3>way from k<BACKSPACE1>the café, <BACKSPACE2> which <2.792><BACKSPACE8>. <1.505><MOUSECLICK><2.931>**After that walked away** and <BACKSPACE5>.

2.3. The Spoken Data: Data Collection Methods

The spoken data were collected through the software Audacity version 3.0.2 (Audacity 2021), and the audio files subsequently were transcribed using CHAT, an established transcription format for corpora (MacWhinney 2000). The purpose of the transcriptions was to annotate such disfluency phenomena that previous research has associated with planning processes and increased cognitive load: pauses, fillers (*ehm*), word fragments, self-corrections, and repeated words (cf. Clark and Wasow 1998).

First, the transcriptions include indications of repetitions of words and word fragments. The transcriptions were carried out using standard Swedish orthography (SAOL 2015) in line with the purpose of the research (Norrby 2014): we wanted the possibility to investigate the content of the message and had no purpose of analyzing it phonetically. However, deviations from standard orthography were made for some common function words. These may be pronounced more "written-like" in stressed contexts, which may be associated with more time for thinking (Johansson et al. 2001; Fox Tree and Clark 1997). More specifically, the conjunction "och" (*and*) can be pronounced either as a short /å/ or /och/ depending on the context. In Table 2, there are instances of "och" being pronounced as both /och/ and /å/. Further, the infinitive marker "att" (*to*) can be pronounced as /å/ or /att/. When participants have used lexical items typically associated with spoken varieties, this has

been included in the transcription. This applies to Alfa's adding of a vowel to the word "här" (*here*) so that it becomes /härа/, making the word longer. It can also apply to forms such as "nån" (short for "någon", somebody) or "sån" (short for "sådan", such).

Table 2. Spoken data An example of the transcription of a spoken truthful account. Periods within parentheses (.) denote silent pauses that are 200 ms or longer, *&-eh* denotes filled pauses, [/] denotes exact repetitions with angular brackets around the preceding strings to show what is repeated, [//] denotes repetitions with minor reformulations, and [///] denotes reformulations. Verbatim repetition is highlighted in boldface.

Transcription of Spoken Truthful Account from Alfa	English Translation
&-eh och &-eh den härа tjejen då får [//] &-eh börjar plocka ihop sin väska &-eh (.) och &-eh sätter sig igen (.) &-eh den härа mannen då han börjar plocka i ordning va han vill ha (.) &-eh och &-eh (.) &-eh den härа tjejen då (.) som hennes väska trillade ner hon [///] &-eh när mannen är och &-eh plockar på ett annat bord så går hon fram å häller ner peppar <**i hans**> [/] &-eh **i hans** mugg	&-eh and &-eh this here girl then gets [//] &-eh starts to pack up her bag &-eh (.) and &-eh sits down again (.) &-eh this here man then he starts to pick up what he wants (.) &-eh and &-eh (.) &-eh this here girl then (.) whose her bag fell down she [///] when the guy is and &-eh picking at another table then she walks up and pours pepper <**in his**> [/] &-eh **in his** mug

Second, the transcriptions contain annotations of filled and silent pauses. To define pauses, we used a common solution from the field of Conversation Analysis (CA), that is, to define a pause from the listener's perspective, in other words, a pause will be defined as a perceived length of silence (following Sacks et al. 1974). However, we also adopted a minimal length of 200 ms. These silences are denoted by (.) following the CHAT format. Filled pauses—i.e., instances of the speaker filling the silence with a sound such as *eh* or *um*—are, according to the transcription standard, denoted by *&-eh*. For the purposes of this study, we used a common transcription standard for all *eh*-sounds and did not discriminate between variations in pronunciation.

Further, other disfluencies in speech, such as repetitions and reformulations were denoted by square brackets and one to three forward slashes, according to the CHAT format. One forward slash, as can be seen towards the end of the excerpt in Table 2, denotes an exact repetition (in this case, "i hans [/] &-eh i hans") and angular brackets denote which words are repeated (if more than one). Two forward slashes denote repetitions with small reformulations, such as at the beginning of the excerpt in Table 2, where Alfa says "då får [//] &-eh börjar" (*then gets [//] &-eh starts*), and three forward slashes denote larger reformulations, such as in the middle of the sample, where Alfa says "hon [///] &-eh när mannen är och &-eh plockar på ett annat bord så går hon" (*she [///] when the guy is and &-eh picking at another table then she*). We will refer to translated excerpts in our examples in this study (see the Supplementary Materials for full versions of the transcripts of the spoken texts).

3. Results

As is evident from the outline of spoken and written language production above, there are many similarities between the two modalities: both require that speakers/writers plan the content and form of the utterance, execute this as a linguistic expression, and evaluate the result against the plan. However, due to inherent differences between the conditions for the modalities, the signs of increased cognitive load will be manifested differently.

The Results section illustrates various incidents deemed suitable for exploring the distribution of cognitive load throughout narrative accounts and outlines methodological issues concerning the definitions and choices of measures used to capture cognitive load in previous language production research. The presentation of results is structured as follows: first, an examination of meaningful measurements for text length; second, an exploration of

the definition and analysis of pauses; third, an investigation into aspects of revisions; and fourth, a proposition suggesting that so-called fluency measures, which integrate all these aspects, may effectively discern segments of increased cognitive load during language production. What is discussed here are thus data from spoken real-time discourse, captured by detailed transcripts, and linear written texts that illustrate the step-by-step actions that writers engage in during text production. However, this article is not concerned with measures for exploring static, final texts. All measures are exemplified and related through excerpts from Alfa and Bravo's accounts.

3.1. Measuring Text Length in Spoken and Written Text Production

Previous studies of truthful and deceitful accounts have highlighted the importance of considering the text length (Knapp et al. 1974; Newman et al. 2003; Derrick et al. 2013), and linguistic comparison of text length in different speakers/writers further emphasizes that both individual differences (for example, age, education, and linguistic proficiency), genre differences, and spoken and written differences influence the text length (cf. Biber 1988; Johansson 2009). Following this, a suitable starting point in describing cognitive load during text production, independent of modality, is to estimate the amount of language production, i.e., the quantity produced by speakers/writers. There are several reasons for this. One assumption posits that the ability to produce longer texts may reflect a less cognitively demanding production process, indicating reduced cognitive load. Another assumption suggests that longer texts may indicate more changes, additions, and explanations, potentially enhancing the credibility of a fabricated narrative (cf. Undeutsch 1989). Measuring text length also serves as a foundational baseline metric for other relevant measures (such as pauses and text changes; see below). Thus, measuring text length will generate an overview of how easy it may have been to re-tell the events of the narrative and will also perhaps render a rough estimation of how elaborate the story is.

The most commonly proposed unit for measuring text length in studies including both spoken and written data has been the *word*, which has also been one of the most important units for measuring text length in deception studies (e.g., see Vrij et al. 2008; Colwell et al. 2007; Derrick et al. 2013). This includes studies illustrating linguistic development (e.g., Berman and Verhoeven 2002) or broad approaches to genre differences (Biber 1988). Importantly, these studies have compared final written texts to transcripts of speaking, i.e., a product to a process (although the transcriptions in these studies can vary according to the degree to which they account for e.g., pauses and repetitions). The reason for this choice is that writing processes have rarely been captured, making studies with process-level descriptions of spoken and written discourse sparse.

Text length in spoken and written discourse has also been compared based on syntax (that is, clauses, sentences, and sentence-like structures). In such comparisons, one major challenge has been that spoken language often lacks the written-like type of grammatical sentences. Therefore, a measure called the T-unit (Terminal Unit; Hunt 1970), which is defined as "[o]ne main clause plus any subordinate clause or non-clausal structure that is attached to or embedded in it," was introduced. It is a syntactic entity, and it is roughly equivalent to a "sentence" in written language. A T-unit is not only defined regarding its syntactic information, but it is also possible to use clues from intonation and discourse/thematic content. As such, it has proven useful for describing and understanding syntactic structure and grammatical development in speech and then comparing it to writing (see Berman and Verhoeven 2002; Johansson 2009; Scott 1988). An addition to the T-unit is the C-unit (*Communication Unit*), proposed by Loban (1976), which allows for utterances without clausal structure to be organized syntactically (see Johansson 2009, p. 93 for an expanded discussion). However, while these measures have been used to compare spoken and written language, the comparison is based on *dynamic* speech and the final products of writing—thus, these measures share the same problems as using the word as a measure. However, there have been attempts to apply T-units to the dynamic linear files from keystroke logging. One example is found in Bowen (2019), where some actions of

revision were removed from the keystroke logging files to better fit the T-unit analysis. This may illustrate the difficulties in applying T-units to linear files. Finally, while the T-unit/C-unit has much in common with "the sentence", it is, in many cases, not equivalent to a graphical sentence that starts with a capital letter and ends with a full stop. The reason for this is that a written, graphical sentence may consist of several main clauses. Thus, investigations using T-units require manual coding of the data. Although it currently seems challenging to apply the notion of T-unit or C-unit to the real-time written data collected by keystroke logging due to the manual work that is needed and the difficulty in identifying T-units in linear text files, it may, for some purposes, be fruitful to explore this measure in the future if one is looking for a rewarding way to compare spoken and written real-time data on a syntactic level. To summarize, while comparisons of text length from a syntactic perspective may prove rewarding, especially through the use of T-units, this is mainly a measure that is suitable for exploring final written texts.

From the discussion above, we can conclude that, in studying real-time data of spoken and written processes, it is complicated to use measures that are adapted for examining final texts. Let us, however, explore the options to use "the word" as a unit a bit further. The traditional definition of a word is a string of letters surrounded by spaces in printing, corresponding to a distinct unit with meaning; however, this definition is difficult to apply to multi-word units, such as "in spite of". For example, in Table 3, the sample of the spoken truthful account of Alfa in the left column includes several occasions of filled pauses (*&-eh*). This raises the question of whether a filled pause should be included in a word count. Another question concerns the instances when Alfa rephrases the information about a bag that is turned over, as in "och välter ner (.) &-eh den hära tjejen som satt på en stol hennes väska &-eh välter omkull den" (*turns over (.) &-eh this girl who sat on a chair her bag &-eh turns it over*). Should both versions of how the bag was overturned be included in the word count? The important point is that there is no correct answer here; instead, the choice of inclusion or exclusion depends on the purpose of the study.

Table 3. Spoken versus written data One spoken and one written extract from Alfa's and Bravo's truthful accounts with different measures and calculations for text length and pauses. All (.) denote pauses and &-eh filled pauses.

Spoken truthful account from Alfa	**English translation**
&-eh när hon har bestämt sig så går hon och &-eh betalar (.) &-eh å då kommer även en man in (.) &-eh ganska (.) &-eh raskt och välter ner (.) &-eh den hära tjejen som satt på en stol hennes väska &-eh välter omkull den så den hamnar på marken (.) &-eh han märker inte de	&-eh when she has made up her mind she goes and &-eh pays (.) &-eh and then enters a man (.) &-eh pretty (.) &-eh quickly and turns over (.) &-eh this girl who sat on a chair her bag &-eh turns it over so it falls on the ground (.) &-eh he does not notice that
Written truthful account from Bravo	**English translation**
Han välde <BACKSPACE3>**te** hennes väska och alla henens <BACKSPACE4>**nes** grejer åkte ur. <2.809>Istället för att säga förlåt **för**<BACKSPACE2>**ortsatte** han att gå <2.183>till **akf**<BACKSPACE3>kaffebordet och **r**<BACKSPACE1>trängde sig framför den unga <1.388>tjejen och tog kaffet förre<BACKSPACE2>**e** henne.	He turnet <BACKSPACE3>ed her bag and all hres <BACKSPACE4>er stuff fell out. <2.809>Instead of apologizing cun <BACKSPACE2>ontintued he to walk <2.183>to the foc <BACKSPACE3>coffee table and r<BACKSPACE1>squeezed in front of the young <1.388>girl and took the coffee beforre <BACKSPACE2>e her.

Regarding the sample of a written truthful account from Bravo, new challenges arise (see the bolded fragments in Table 3). Corrections of misspellings here result in word fragments. For instance, the word "välde" (*turnet*, a misspelled version of *turned*) involves three presses on the backspace to delete one space and the two last letters. The correct letters "te" are then immediately added to create "välte" (*turned*). How should we calculate words in this context? Should "välde" be considered one word and "te" another, should

we treat "välde" and "välte" as two separate words, or should all be counted as one final, correct word, "välte"? From the perspective of studying deceitful narratives, one could argue that the correction of misspelled words is uninteresting, as such corrections merely contribute to the surface level of the final text. However, it may influence the cognitive load in a way that writers focusing on low-level orthographic processes have fewer cognitive resources available for other activities. With this argumentation, it is essential to establish a method for including word fragments and alternative spelling varieties as they contribute to the understanding of how the writers' resources were distributed during their writing.

In summary, spoken texts include word repetitions, and a decision to count each repetition of the same word (or phrase, or part of the phrase) leads to the risk of obscuring parts of the process only once. Issues also arise regarding whether "filled pauses", (e.g., *um*, *eh*) should be counted as words or not. Yet another issue is word fragments, which occasionally occur in speaking, that is, the instantiation of words where only a few speech sounds are included and when it is sometimes difficult to guess which word was intended.

In writing, on the other hand, fragments are frequent, and just as in speaking, the mapping of the fragments into words can pose challenges. The fragments may comprise just one or two letters, where it is impossible to comprehend whether they are signs of a false start or mistyping. However, often the fragments present alternatively spelled versions of the same word, and it is common for only a portion of the word to be deleted and rewritten. Should these instances be calculated as one word or two words (or more)? Furthermore, it is typical for writing to encounter letter combinations that arise accidentally or erroneously by pressing the wrong key. Finally, in writing, it is not uncommon for entire clauses, sentences, or paragraphs to be deleted, rewritten, or pasted and moved around within the text. Just as in spoken language, a decision must be made regarding whether to count the words and phrases once or multiple times.

When comparing the text length of written texts, for instance, across genres or between or within subjects, a proposition is to use the number of written characters (i.e., letters, numbers, punctuations, and spaces in writing) in the linear text (see example in Johansson 2009). Such an approach would account for fragments, words, phrases, and other parts of the texts, which may have been deleted and are thus invisible in the final text but which are nevertheless part of the work that writers have put into producing a text.

Spoken texts do not offer the same possibility to easily capture all phonemes. However, with a carefully conducted transcription of the spoken process, the result will not only be a written, linear reflection of the spoken process, but it can also serve to capture text length as the number of written characters in the transcriptions. Although this is not the same as a phonetic transcription that encompasses all phonemes, this approach will serve the purpose of estimating text length in relation to cognitive load and, importantly, enable a rough estimation of how truthful and deceitful accounts compare within and across modalities.

Table 4 illustrates the outcome of the different ways of measuring text length: number of words and number of characters. In this example, the computations are conducted using the excerpts in Table 3. Here, it is evident that variations exist in the word count of the written linear texts that are contingent upon the inclusion or exclusion of word fragments.

Hence, our optimal recommendation for comprehensively understanding the text length of spoken and written accounts while concurrently capturing the expression, repetition, or reformulation of small units is by measuring the number of characters in writing and the number of characters in the *transcription* of spoken accounts. In doing so, we propose that, for some purposes, it may be suitable to include filled pauses in speech to account for the fact the speaker is uttering something, in contrast to being silent. In our methodological approach, a filled pause denoted as *&-eh* in the transcription would be computed as two characters (*eh*). Silent pauses in speaking would be excluded from the calculation, just like the pauses in writing.

Table 5 provides an overview of the number of characters, with filled pauses in speaking being annotated and counted as two characters (i.e., *eh*; see Section 2.2 for an elaboration

of the transcription decisions). See the Supplementary Materials for the complete accounts for both Alfa and Bravo.

Table 4. Text length measures Calculations of the number of words, characters, and pauses and the number of characters divided by silent, filled, and all pauses (this metric thus delineates the number of characters produced between pauses). The numerical values are derived from the excerpts outlined in Table 3. The initial column enumerates the variables under consideration, the second column delineates the calculations for Alfa's spoken excerpt, and the third column expounds the calculations for Bravo's written excerpt.

		Spoken	Transcriptions	Written Linear
Number of words		52	*With fragments*	39
			Without fragments	33
Number of characters		237		203
Number of pauses	*Silent*	5	*All pauses*	3
	Filled	8		
	All pauses	13		
Characters per pause	*Silent*	47.4	*All pauses*	67.67
	Filled	29.63		
	All pauses	18.23		

Table 5. Descriptive statistics for language production across the four accounts. Alfa's spoken and Bravo's written. Time on task represents how many seconds Alfa and Bravo spent speaking/writing. Number of characters represents the total number of characters in the different accounts. Characters per second represents the average number of characters produced per second. The proportion of deleted characters demonstrates the proportion of the written text that was deleted. Number of pauses represents the total number of pauses. Revisions and reformulations represent the number of editing operations independent of editing size.

		Spoken Account from Alfa			Written Account from Bravo	
		Truthful	**Deceitful**		**Truthful**	**Deceitful**
Time on task(s)		127.441	128.78		353.93	259.695
Number of characters	*Including filled pauses*	1382	1468	*Final text*	1306	1208
	Excluding filled pauses	1268	1373	*Linear text*	1521	1310
				Number of deleted characters	215	102
				Proportion of deleted characters %	16.5	7.8
Characters per second	*Including filled pauses*	10.844	11.399	*Final text*	3690	4652
	Excluding filled pauses	9950	10.662	*Linear text*	4297	5044
Number of pauses	*Total pauses*	68	60	*<2 s*	33	23
	Silent pauses	30	29	*Total pause time (s)*	100.973	65.906
	Filled pauses	38	31	*Mean pause duration*	3.06	2.865
Revisions and reformulations		4	7		51	24

3.2. Defining and Measuring Pauses in Spoken and Written Text Production

Regarding the spoken account in Table 6, numerous examples of silent (.) and filled (*&-eh*) pauses are discernable in the transcription. Similarly, in the written account, examples of pauses (<2.809>) are evident in the linear representation of the writing process. As detailed earlier, pauses during language production are closely linked to moments of increased cognitive load. Consequently, the identification of pauses and their location and duration are highly relevant for our purposes.

Table 6. Pauses in spoken and written accounts. Examples illustrating the processes of linguistic changes in the spoken deceitful text by Alfa and the written deceitful text by Bravo. The (.) denote pauses and &-eh filled pauses. [//] denote repetitions with reformulations.

Transcription of spoken deceitful account from Alfa	English translation
och &-eh (.) &-eh personen till vänster &-eh (.) &-eh (.) tar fram en fusklapp ur fickan (.) och &-eh (.) försöker lite diskret att skicka över den till &-eh personen till höger (.) &-eh (.) den här personen till höger vill [//] försöker säga nej å vill inte ta emot lappen men tar till slut emot den &-eh och &-eh <tanken är att> [//] (.) eller hon [//] man ser att hon försöker &-eh (.) lämna tillbaka den men &-eh hon hinner inte för tentavakten &-eh har precis torkat klart på golvet	and &-eh (.) &-eh the person to the left &-eh (.) &-eh (.) takes a cheat note from the pocket (.) and &-eh (.) tries little discretely to send it over to &-eh th person to the right (.) &-eh (.) this person to the right wants [//] tries to say no and does not want to take the note but takes in the end it &-eh and &-eh <the idea is to> [//] (.) or she [//] one sees that she tries &-eh (.) to return is but &-eh she has no time because the exam teacher &-eh has just finished wiping on the floor
Linear written deceitful account from Bravo	**English translation**
Samtidigt som <1.302>han <2.292>har ryggen vänd mod <BACKSPACE2>t std<BACKSPACE1>udenterna tar tjejen med gul tröa<BACKSPACE1>ja upp en lapp ur sin ficka. Hon <3.466>vecklar snabbt ut lappen och <1.779>läser <1.196>kort innan hon knägglar <BACKSPACE7>ögglar ihop den igen. <2.812>När <2.079>tentavakten vänder sig om igen är lappen borta och flick<BACKSPACE5>tjejerna fortsätter att skriva.	Same time as <1.302>he <2.292> has (his) back turned tii<BACKSPACE2>to std<BACKSPACE1>udents the girl with yellow shirt taks<BACKSPACE1>es up a note from her pocket. She <3.466>unfolds the note quickly and <1.779>reads <1.196>short before she crankles<BACKSPACE7>inkles it together again. <2.812>When <2.079> exam teacher turns back again the note is gone and the (young) gir <BACKSPACE5>girls continue to write.

So, what is a pause? Pausing during writing (on a keyboard) is often defined as inactivity between two keypresses. However, it is debatable how long the duration of the inactivity should be for it to count as a pause. Technically, pauses can be defined as short as the hardware accounts for, that is, there is generally a latency between the pressing of a key until it is registered. In writing theories (e.g., Flower and Hayes 1981), longer pauses are highly associated with high-level processes such as planning processes (see Torrance 2016; Torrance et al. 2016), while shorter pauses, to a greater extent, have been seen as indicative of low-level processes related to transcription, orthography, and spelling (Wengelin 2006). The literature does not propose strict cut-off points when a pause is long or short, and instead, this must be seen as relative given the particular task or circumstance. However, in general, writing researchers adopt ad hoc pause criteria based on the purpose of the study. Often pauses of 2 s and longer have been proposed as indicating high-level processes (Wengelin 2006), while shorter pauses have been associated with the low-level processes.

The variation of pauses between subjects has been acknowledged in many studies (see Spelman Miller 2006b; Lindgren et al. 2019) and has been explained through background factors (such as writing in first or second language, education, and practice in writing (including handwriting and/or typing skills), age, linguistic development, and reading and writing difficulties due to dyslexia or aphasia), and contextual factors (topic and genre knowledge, audience awareness, or occasional disturbance in the surroundings). Further attempts have been made to propose methods for establishing individual pause criteria, which would allow for a more reliable comparison between subjects. Proposals include correlating the individual pausing behavior to writing speed (Chenu et al. 2014), or relating to the dynamic variation of keypresses across a writing session (Olive 2014). Thus, writing studies examining pauses must establish their own ad hoc criteria for how to define a pause and how to discriminate between long and short pauses if that is relevant given the research questions, and they must use an experimental design that controls for within- and

between-subject factors. To facilitate the analysis of pauses, we have used the ad hoc criteria of 1 s. This will allow us to capture pauses on a relative micro-level but avoid having to address pauses that may be primarily related to transcription skills (see Wengelin 2006).

Regarding speaking, the definition of a pause is equally tricky, not the least the issue of individual variation that applies to this modality as well. When (silent) pauses are investigated in speech, there has been a general acknowledgement that the definition of a pause must be related to individual speaking rates. However, the speaking rate may vary between sessions and within the same session. A common solution in the CA transcriptions is to include so-called perceived pauses (Sacks et al. 1974), which is how we operationalize it (while only including pauses with a minimal length of 200 ms).

In addition, a decision on whether to treat filled and silent pauses equally or not must be made. Do filled pauses (*eh*, *um*, etc.) serve the same purpose as silent pauses? Studies on conversation show that filled pauses are communicative and can help the speaker keep their turn, whereas silent pauses are not necessarily so (Clark 1996). However, while our data contain spoken *monologues* where "keeping the floor" should not be an issue for the speakers, there are still ample examples of filled pauses in the data. For instance, in Table 4, it is shown that, of a total of 13 pauses in the spoken sample, 8 are filled. Occurrences of filled pauses in monologues should perhaps be interpreted along the lines that speakers have incorporated filled pauses as one of several planning strategies during speaking and that it is difficult to abandon this habit when invited to speak uninterrupted.

To sum up, pauses are seen as important indicators of speakers' and writers' increased cognitive load during language production. However, it can be difficult to define a pause; previous research has established a rough standard for the respective modality, and we have employed these standards in our analysis. Since there may be different preferences for using filled or silent pauses, which may vary between and within different accounts of speakers, we measured the number of silent pauses as well as the number of filled pauses. In addition, when appropriate, we advocate a measure where both types are included—for instance, to illustrate the overall number of pauses.

Once we have established the definition of pauses in each modality, we turn to measuring *the number of pauses*. One can assume (based on, e.g., findings from Goldman Eisler 1968; Heldner and Edlund 2010) that frequent pausing would be an indication of instances where the cognitive load is increased and where the speakers/writers need extra time to think about the linguistic expression. With our definition of pauses, it is relatively easy to calculate the number of pauses in a written or spoken account from the linear files in writing or the transcription of the spoken accounts. Since the text length and/or the amount of time dedicated to the accounts will differ, the number of pauses must be calculated relative to text length and/or writing/speaking time.

Further, *pause duration* has been proposed as an important indicator of cognitive load, and longer pauses are often found preceding more linguistically complex constructions, e.g., subordinated clauses or complicated noun phrases in both speaking (Goldman Eisler 1968) and writing (Nottbusch 2010). While there are tools that can identify silences, these will be rendered useless if there is any kind of background noise in the audio file and filled pauses will also not be captured with these tools. Thus, in speaking, the calculation of pause length will need manual attention and consequently be very time-consuming. Written data, collected through keystroke logging, will have various options for calculating pause length readily and will be automatically accessible (Leijten and Van Waes 2013).

Pause location is a final component that is likely to be relevant for addressing cognitive load during language production. Pause location in connection with specific syntactic constructions, or semantic information may reflect, on the one hand, difficulties in structuring the message, or, on the other hand, difficulties with finding lexical expressions that reflect what one needs to say (Matsuhashi 1981; Spelman Miller 2006a). These types of investigations may be rewarding in establishing which segments are particularly challenging for speakers/writers from a forensic perspective. Pause location can, to a certain degree, be annotated in the transcriptions with the use of speech technology tools, such as linguistic

parsers that indicate parts of speech (there are a few parsers trained for Swedish, e.g., Qi et al. (2020), which could aid in this). However, one must expect that substantial manual handling is needed, not the least since transcriptions with pauses and repetitions will make automatic analyses difficult.

3.3. Revisions and Reformulations in Spoken and Written Language Production

This section addresses, on the one hand, revisions in writing and how they may be expressed and studied, and, on the other hand, how reformulations and repetitions can be studied in speaking. We treat all these aspects as manifestations of changes in the linguistic message. According to the models of writing and speaking, such changes would occur by monitoring what has been previously produced and will happen if the speakers or writers after such an evaluation conclude that the previous text needs to be modified.

We start by outlining how the concept of revision has been described in writing. Its complexity is discussed in a seminal article by Faigley and Witte (1981), where they make the point that revision should not only be viewed as tidying up the text after the first draft. Instead, there is substantial evidence of it being a complex process that writers engage in, concurrent with planning new content and generating text. Therefore, reading or monitoring the text written so far is an important component (see Johansson et al. 2010; see Wengelin et al. 2023). Changes in the written text can be made at any point during the composition of text: before the text has been transcribed, at the point of inscription (i.e., at the end, or *the leading edge*, of the text being produced), or at a previous point in the text (cf. Fitzgerald 1987; Lindgren et al. 2019).

There are undoubtedly revision processes of different kinds, and from a processing point of view, there can be so-called *internal revisions*, also referred to as pre-linguistic and pre-textual revisions (see Murray 1978), which will occur mentally and never be manifested or overtly expressed. *External revisions*, on the other hand, can be made at the point of inscription or in the previous text. Revisions can further be classified as *surface revisions*, i.e., language revisions (associated with formal changes), or *deep revisions*, i.e., content revisions (associated with semantic information) (Chanquoy 2009; Stevenson et al. 2006). The concept of internal revision can further be compared to the idea of text generation as part of the planning process in the model of Flower and Hayes (1981). It is important for the purpose of our exploration that some revision processes may not be overtly visible in the written data, but instead, to a certain extent, incorporated in pauses where the writer is planning the linguistic expression and trying out and rejecting possible solutions before settling on one decision. Existing literature provides many examples of different taxonomies for categorizing the types of revisions occurring in writing, where adding, deleting, and substituting content are the most agreed upon (for some examples, see Johansson et al. 2023).

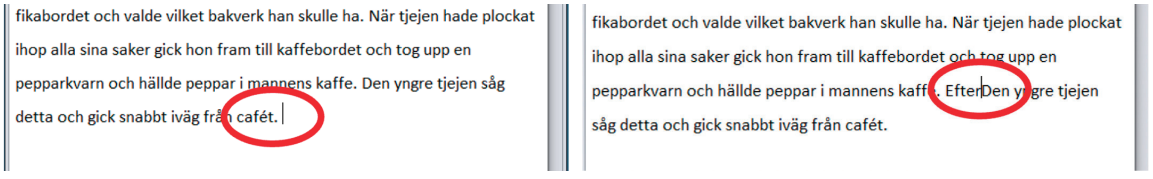
Just like other linguistic processes, the acts of revision will fluctuate depending on the context, the task at hand, and the background of the writer. Here, age, education, linguistic proficiency (writing in the first or second language and grammatical and lexical knowledge), writing proficiency, knowledge of the topic and genre, and writing mode (for example, typing and handwriting) will influence how, what, and when revisions occur (for overviews, see Chanquoy 2009; Lindgren 2005).

Table 6 provides examples of Bravo's revisions in the deceitful written account, mostly consisting of surface revisions at the leading edge, where typos (e.g., errors occurring due to the pressing of the wrong key and not because of ignorance of orthography) are corrected. The erroneous 'd' at the end of *mod* is changed to *mot* ('towards'); the initial letter combination *std* is immediately corrected and the word *studenterna* ('the students') is written; the misspelled word *tröa* is at once corrected to *tröja* ('sweater'). One change can be categorized as a content revision, where *flick (orna)* ('young girls') is changed mid-word to the (near) synonym *tjejerna* ('girls'). Similar surface revisions at the leading edge are found in Table 3, in the linear text of Bravo's truthful written account. Here, we see no examples that can be categorized as content revisions. The examples of revision in these

excerpts thus show how the writers immediately tend to surface revisions (note that there are no pauses between the deleted written text and the use of backspace and the added written text), which suggests a constant monitoring of what is being written. We have also seen examples of content or semantic revision, where another lexical choice for "the girls" was made.

The linear files of the writing sessions further allow for the study of other types of revision behavior: using the arrow keys or mouse to move around in the text. Such movements may or may not be followed by a backspace (for deleting text) or the addition of text to previously written parts. Writers can also highlight parts of the text by using the mouse and click and drag functions, or by using combinations of keys (shift + alt and arrow keys). Once highlighted, the text can be deleted, moved, copied and pasted, or overwritten if writers type over the highlighted text with new text. Consequently, a lot of text can be deleted or moved with very few keypresses or mouse movements. Therefore, it can be relevant to account for the number of *editing operations* that take place independent of how much text is being removed or added in each operation. These types of editing operations are unique for keyboard writing, but the same concept can be adapted for speaking if reformulations or self-repairs are included in calculations. In Table 5, the total number of editing operations for the complete sample files (found among the Supplementary Materials) is included. The numbers in the table further illustrate how common editing operations are in writing, compared to reformulations in speaking.

The full writing session of the truthful account by Bravo can provide an illustration of what it can look like when a revision is made away from the leading edge, in the previously written text. By the end of the final text of Bravo's truthful account (see the Supplementary Materials), she uses the mouse to move the cursor to a spot preceding the last written sentence. There, she adds a sentence. Table 1 shows the linear file of this sequence, where the indication of <MOUSECLICK> is seen on the last line, followed by a pause of 2.931 s, and then, the sentence fragment that was added (in boldface in Table 1): "Efter det gick ut därifrån och" (*After that went out and*). Note that this sentence lacks a subject, possibly the pronoun "she", and that the last word "och" was immediately deleted. An illustration of what it looks like is found in Figure 1, where we see two screenshots from the real-time replay of the writing session: the first one just before the mouse click and the second one immediately after the first word of the new sentence has been written ("Efter", *After*). In the figure, the red circles show the placement of the cursor in the two examples.



(**a**) (**b**)

Figure 1. Text revision away from the leading edge. Example of inscription points from Bravo's truthful written account. (**a**) Bravo writing at the leading edge; (**b**) Bravo has moved the cursor and is now inserting a sentence away from the leading edge. The red circles denote the placement of the cursor.

As mentioned above, revisions can occur far from the inscription point, that is, when the writer uses the mouse or the arrow keys to move the cursor away from the inscription point to change something that has already been written (Lindgren et al. 2019). This means that writers can add, delete, or change the previously written text at any point during the writing session anywhere in the text. For example, a writer may add an initial paragraph of the text as the last part of the writing process, change a description of a protagonist, or delete a chain of events. In the final text, there will be no trace of this (see Wengelin

et al. 2023 for examples of this execution in advanced writers). However, examining when writers decide to make changes in their previous written texts offers new perspectives for the understanding of how the message is constructed and can give insights into how deception is built. One example of what revisions may look like when they occur away from the leading edge is shown in Figure 1, where the writer Bravo has finished a sentence ("Den yngre tjejen såg detta och gick snabbt iväg från cafét" *The younger girl saw this and quickly departed from the café*) and then, she moves the cursor to before this sentence to add a new sentence ("Efter det gick [de] ut därifrån" *After that they left*.) These kinds of revisions do not have an obvious equivalence in speaking, which probably can be attributed to changes in speech due to the necessity for immediate changes—using the terminology from revision in writing, one can say that changes during speaking will occur in a linear fashion and always at the leading edge. It is undoubtedly an option for speakers to address something that was said further back in the spoken message, and draw the attention to what they need to change or add information to what was previously stated. However, speakers can never "move away" from the leading edge of their spoken account. The phenomena of "revision" in speech are typically referred to as *disfluencies* in the literature (see Clark and Wasow 1998). This term covers filled and silent pauses, prolongations, repetitions of words and utterances, as well as reformulations. The psycholinguistic view on disfluencies is expressed in Goldman Eisler's (1968) seminal work that connects increased number and duration of pauses and other signs of disfluencies with increased linguistic complexity (especially regarding syntactic complexity at the clause or phrase level). Similar views are shared by Clark (1996) and Levelt (1989) (see also Eklund 2004 for an overview, with a phonetic focus on disfluencies in speech). Here, we will mainly be concerned with repetitions and reformulations since they, just like revisions in writing, serve the purpose of being overt changes to the linguistic message.

A common example of disfluencies in speaking is to repeat one or more words occurring at the start of a clause verbatim, a strategy that is often associated with planning (Clark and Wasow 1998). Table 2 shows an illustration of this from Alfa's spoken truthful account (verbatim repetition in boldface). She says "när mannen är och &-eh plockar på ett annat bord så går hon fram å häller ner peppar <**i hans**> [/] &-eh **i hans** mugg" (*when the guy is and &-eh picking at another table then she walks up and pours pepper <**in his**> [/] &-eh **in his** mug*). Here, the repetition (*in his*) occurs at the end of the clause, where it precedes the noun "mug". Note that in connection with the repetition, we also find a filled pause (*&-eh*).

Table 6 illustrates parts of the spoken deceitful account from Alfa, which contains numerous examples of reformulations. She says "personen till höger" (*the person on the right*), which is followed by a silent pause, a filled pause (*&-eh*), and another silent pause. She then says "den här personen till höger vill" (*this person on the right wants*), and then, the last verb ("vill") is changed to "försöker" (*tries*). Thus, taken together, this is a sequence of self-repair consisting of a series of reformulations of what is pretty much the same content. Just a little bit further on, Alfa has another sequence of reformulations: "tanken är att" (*the idea is to*), which is followed by a pause and the fragment "eller hon" (*or she*), which, again, is abandoned for the clause "man ser att hon försöker" (*one sees that she tries*). First, these kinds of sequences of reformulations are particularly interesting to study because they highlight a circumstance or event that the participant finds difficult to express in words. Second, they constitute a noteworthy example of how the strategy of "talking around" a subject allows more time to think while at the same time ensuring no interruptions from listeners—a purpose often attributed to filled pauses. Finally, this example illustrates a sequence of (extensive) consecutive revisions. For our purposes, such sequences are intriguing in both modalities as they have the potential to reveal particularly challenging portions of the narrative accounts.

These instances of verbatim repetition and consecutive reformulations during speech demonstrate that speakers often make repeated attempts to find the right expression with the rephrasing frequently involving multiple words. Notably, the observed changes in our

examples appear to be more closely associated with linguistic content, specifically lexical choices, rather than linguistic form.

For our objectives, it is pertinent to explore methods of quantifying revisions, repetitions, and reformulations as a cumulative display of such occurrences may indicate disturbances in the planning processes due to heightened cognitive load. One approach, used in writing studies employing keystroke logging technology, involves subtracting the final text's character count from the character count in the linear files (see example in Gärdenfors and Johansson 2023). This will result in a proportion of the text that was deleted (see Table 5 for an example from our data). Another option is to calculate how many *editing operations* there are (i.e., the number of occasions something was deleted, independent of how much text was deleted each time, cf. Johansson 2000). In both cases, this will demonstrate a quantitative approach to capturing how frequent revision occurs.

In spoken language, the concept of "deleted text" becomes irrelevant as all utterances, whether rephrased or not, are overtly expressed. However, quantifying and accounting for the number of repetitions and reformulations provides an overview of how frequently speakers rephrase themselves.

An additional potentially valuable approach to investigating changes would be to annotate their location or context and/or categorize the nature of the revision. This could shed light on the causes of increased cognitive load, following the insights of Goldman Eisler (1968), and reveal whether the linguistic expression leading up to deceitful information is more prone to revision or if the deceitful information itself is the focus. However, it is important to note that such annotations necessitate manual execution, making it a time-consuming task.

3.4. Fluency and Disfluency in Spoken and Written Language Production

We have touched upon that accumulative signs of cognitive load may be of relevance for our purposes—that is, where pauses and/or changes occur together or within a short time frame. In addressing this issue, we turn to the concept of *fluency–disfluency*. For speaking, the concept of fluency has been an important concept for estimating how easily speakers carry out different oral tasks. There are many examples that comprise proficiency in second-language learning (e.g., Jong 2016) or fluency in regard to disturbances during speaking, for example, stuttering (e.g., Alm 2011). In the study of spontaneous speech (whether from a cognitive approach or CA perspective), it is also contrasted against the notion of disfluency, which would be viewed as unwanted disturbances during speaking (Clark and Wasow 1998; Eklund 2004; Norrby 2014).

In writing, fluency was brought to the forefront by Chenoweth and Hayes (2001) as a way to shed light on linguistic proficiency (often from an L2 perspective, see examples in Manchón and Roca de Larios 2023) and writing competence. Fluency during writing will typically be captured by dividing the number of linguistic units (words or written characters) per time unit (seconds, minutes, or the whole time on task/total writing time) (see Kaufer et al. 1986 for early examples). From a processing perspective, fluency is often measured through "bursts", that is, the number of words or the number of typed characters between pauses (*P-bursts*) or between revisions (*R-bursts*) (see Alves and Limpo 2015 for a comprehensive overview of bursts in writing). Increased fluency will occur when writers have few and/or very short pauses and few revisions. Keystroke logging software, especially the widely used Inputlog (Leijten and Van Waes 2013), can provide automatic output with a variety of different bursts and applied pause criteria. Such output can show the mean length of P-burst in a text, or, in other words, the average number of characters that are written between each pause. According to the hypothesis, the P-bursts will be longer if writers produce new text with ease.

Here, we can refer to Table 4, where the number of pauses in speaking and writing are presented across modalities. The number of pauses is divided by the number of characters. This would be an example of the mean length of a P-burst. For the spoken account, we have included several comparable measures: one measure where characters have been

divided by the number of silent pauses (47.4), one that divides them with the number of filled pauses (29.63), and one that includes all pauses (18.23). The different results illustrate that the definition of pauses is important for the outcome. For the written account, we only included one measure: number of characters per pauses longer than 1 s. Note that, with a different pause criterion, the number of characters per pause would also change. Determining which pause criteria to choose and whether or not to include or exclude filled pauses will depend on the research questions that are posed, but it may also be valuable to explore various options before deciding on the definition in a particular study.

The fluency approach thus requires measuring the total time on task in the written and spoken task. At the overall text level, this would mean that, for the written task, the keystroke logging software offers automatic output, while the spoken task requires some, but limited, manual attention. The number of written characters or the number of characters in the transcripts of the spoken accounts can then be divided by the time on task. We advocate using the linear text files for this type of calculation. In examining the results of such calculations, a few effects can occur: if speakers/writers have long and/or many pauses, there will be, on average, fewer characters written per second. However, if speakers/writers engage in many changes (revisions, repetitions, and reformulations), the effect may be more characters written/spoken per second. It may be the case that writers who have longer pauses also revise more, but not necessarily so. Given the previous studies we have repeatedly referred to above, it is evident to expect a fluctuation regarding where pauses and changes occur during the unfolding of both spoken and written language production.

We have already concluded that it is difficult to automatically identify and isolate pauses in speaking due to potential background noise in the recording and the existence of filled pauses; consequently, we have ruled out measuring pause duration in speaking as a cost-effective way to approach our goals. However, to account for the fluctuation in fluency during language production, we suggest another approach, that is, dividing the spoken and written texts into different segments. Given our experimental design, where we have identified which portions of the events in the elicitation films should be altered by the participants in their deceitful accounts, we propose a threefold division: before the lying event, during the lying event, and after the lying event. This will be a way to operationalize the variation in fluency during different sequences of the narrative accounts and serve as an initial but potentially rewarding attempt that can serve our purposes.

Similar approaches were applied to written data in a study by Johansson (2009), but in that case, the writing time was divided into five equally long segments and then, the proportion pause time was measured in each segment. The results demonstrated different pause time distributions throughout the writing of narrative and expository genres. For our data, this would be a more time-consuming but perhaps fruitful way to divide the speaking and writing into fixed time segments or 20% divisions of the time on task and then explore the proportion of pause time and/or changes in each segment.

Finally, an initial quantitative and cost-effective approach to identifying sequences with accumulated signs of cognitive load can later be combined with more qualitative inspections and annotations. For our purposes, measuring fluency may thus provide an approach that combines text length, pauses, and changes—all outlined above. The advantage of looking at fluency is that it is a proportional measure and thus more suitable for comparing accounts across participants with different text lengths, and consequently, across modalities and deceitful–truthful conditions.

4. Discussion

This study had a methodological aim to explore and identify phenomena indicative of increased cognitive load within language production in both speaking and writing. Our objective was to discern and quantify heightened cognitive load effectively, with the aim of establishing a methodological framework capable of using these effects to identify deceitful narration, which can be an indication of lying. Drawing inspiration from previous research

on cognitive load measures in language production, our study specifically concentrated on indicators within the speaking and writing processes, such as text length, pauses, changes (revisions, repetitions, and reformulations), and fluency.

4.1. Measuring Text Length

The first research question revolved around the methodology for quantifying cognitive load by measuring text length. The examination of text length is pertinent, grounded in the assumption that longer texts reflect ease in both speaking and writing. This characteristic may serve as an indicator for both truthful and deceitful accounts. Moreover, the augmentation of text can signify an individual's effort to enhance the persuasiveness of a narrative. Our deduction from this exploration is that, although words are commonly used as a metric for text length, such a measure may obscure instances of word fragments and revisions in written content.

In speech, it is equally imperative to acknowledge the difficulty in operationalizing repetition and rephrasing of words and phrases. To derive a comprehensive measure for text length that incorporates the entirety of overt linguistic production by speakers and writers, we propose using the character count (including letters, numbers, punctuation signs, and spaces) in written text. This approach encompasses all textual elements produced, irrespective of linguistic unit. For spoken language, an equivalent measure can be achieved by calculating the character count in the transcription of spoken accounts. This methodological choice facilitates the inclusion of word fragments, false starts, repetitions, and reformulations. Additionally, if necessary, the annotations of filled pauses can be incorporated into the calculation.

Despite the inherent limitations of this approach, it represents an ad hoc solution that aligns with the research objectives of the larger project. Moreover, the accessibility of character count information in keystroke logging programs, as well as its ease of extraction from transcriptions of spoken accounts, renders this methodology a cost-effective and economical means of obtaining relevant measures for researchers interested in exploring or comparing spoken and written discourse.

4.2. Measuring Pauses and Pause Length

The second research question delved into the nuanced definition and measurement of pauses to effectively capture heightened cognitive load. Within the literature, pauses are commonly regarded as a robust indicator of increased cognitive load. In addressing this matter, we initiated a discussion on the definition of a pause. In the context of keyboard writing, a pause has traditionally been defined as the interval of inactivity between two keypresses. However, such pauses can be exceedingly brief, making it impractical and often less rewarding, to scrutinize every pause between keypresses. Recognizing this, we advocated for an approach commonly adopted by writing researchers, which involves setting a pause criterion to exclude very short pauses, unless the focus lies on low-level processes such as transcription skills. For the purpose of the larger project, we specifically proposed an ad hoc criterion of 1 s, enabling the capture of pauses on a micro-level relative to the study's scope while circumventing the need to address pauses primarily associated with transcription skills.

In the realm of spoken language, defining pauses poses its own set of challenges. For instance, should one measure each silent pause and only consider pauses surpassing a specific temporal threshold? Drawing insights from conversation analysis, we opted to incorporate the concept of perceived pauses, wherein listeners' perceptions determine what qualifies as a silent pause, but we adapted our definition to include only pauses exceeding a minimal length of 200 ms. This entailed employing slightly different approaches to defining pauses in writing and speaking. Our rationale behind this decision was twofold: to facilitate subsequent analyses related to pauses and to establish criteria suitable for our research needs, grounded in robust practices within the field of speaking and writing.

Additionally, this study addressed filled pauses, such as instances where speakers use fillers (*eh*, *um*) to avoid silence and potential loss of conversational footing. While we deemed it reasonable to occasionally incorporate filled pauses into the same calculations as silent pauses, our data structure in the corpus of the larger project allows for the separate calculation of different pause types when necessary. This flexibility enables comprehensive examination of overall pause distribution during speech and, when desired, discrimination between various pause categories.

Then, we focused on the dimension of pause duration, recognized as a notable candidate for indicating heightened cognitive load. The suggestion was that opting to define pauses as perceived pauses in speaking is preferred, as opposed to the more time-consuming approach of measuring the length of every silent pause above a certain threshold. In contrast, within the written data acquired through keystroke logging, accessibility is straightforward, and various pause criteria can be applied for data exploration. Consequently, although pause length could serve as a vital indicator for capturing heightened cognitive load, its examination demands a more labor-intensive, manual approach.

The location of pauses within the texts additionally conveys insights into the linguistic contexts where writers/speakers allocate additional time. While this approach holds promise, this study ascertained that its implementation necessitates considerable manual effort for the meticulous annotation of syntactic context and semantic content. Notably, the researcher's workload in this task remains equivalent when annotating both spoken and written data following transcription. It is pertinent to acknowledge that leveraging existing parsers for parts of speech tailored for Swedish could assist in this undertaking. However, these parsers encounter challenges when confronted with word fragments, introducing potential limitations and uncertainties in content coding. Despite recent advances in this domain, exemplified by technologies like ChatGPT (https://chat.openai.com/chat, accessed on 26 January 2024) and other AI applications, it is conceivable that novel tools better suited to our needs may emerge in the near future. Nevertheless, the imperative role of manual supervision remains evident to ensure validity. In light of the objectives of the larger project, we conclude that annotating pause location, while potentially valuable from a forensic perspective, does not stand as a primary choice for initially identifying sequences of heightened cognitive load in our data.

4.3. Measuring Linguistic Changes

The third research question centered on overt revisions in writing and reformulations and repetitions in speaking, denoting instances when writers and speakers modify previously produced messages. Presumably, such alterations aim to enhance the message's accuracy or persuasiveness, with an anticipation of increased instances of revision and reformulation occurring in specific linguistic contexts and narrative sequences where expressing the intended meaning proves challenging on the initial attempt. However, the manner in which changes in the linguistic message manifest differs significantly between writing and speaking.

In writing, revision can occur at any point during a writing session, often proximate to the leading edge, and often addressing formal language aspects. However, writers possess the freedom to navigate to any part of the written text, addressing various issues. Typically, minimal traces of revision are discernible in the final text. Leveraging keystroke logging methodology enables the visibility of revisions through linear file inspection and session replay. Supplementary output files can further provide an overview and categorization of revisions (see Leijten and Van Waes 2013).

Conversely, the dynamics of reformulation and repetition in speaking differ. These modifications are both audible and available to listeners. Speakers are confined to making changes at the leading edge and perhaps use the repetition of words and expression as a strategy to gain thinking time in a similar but more sophisticated way as the function of filled pauses is interpreted. Reformulations or self-repairs serve the dual purpose of extending the time for thought, planning (indicating a desire to convey specific thoughts),

and experimenting with different linguistic expressions. Transcriptions of spoken accounts offer a relatively straightforward means of capturing repetitions and reformulations. Although coding is necessary during the initial transcription phase, corpus tools (CLAN (MacWhinney 2000), AntConc (Anthony 2023), etc.) can be used for the almost automatic acquisition of such data.

In summary, regarding research question three, addressing revision in writing and repetition and reformulation in speaking appears to be relatively straightforward. However, if one seeks to annotate the location, specifically the syntactic or semantic context in which the revision occurs, similar challenges as those regarding pause location may arise. Manual coding, especially regarding semantic aspects, becomes imperative.

Yet, the application of keystroke logging software presents a viable solution by offering a string-based analysis of the revision's location in writing, distinguishing between locations such as mid-word, between words, and between clauses (Leijten and Van Waes 2013; Wengelin and Johansson 2023). Consequently, estimating the amount of revision becomes achievable, either by measuring the number of deleted characters in writing—an easily obtainable measure—or by simply counting the instances where a speaker or writer engages in editing operations, irrespective of the size or scope of the revision. The mere occurrence of changes in the text is indicative of potential increased cognitive load.

4.4. Exploring the Fluctuation of Cognitive Load

The fourth research question encompasses an examination of how text length, pause distribution, and text changes, when considered collectively, can facilitate the exploration of cognitive load fluctuations throughout the process of spoken and written language production. From the comprehensive overview of our measures, it is evident that each area holds significant potential for capturing crucial indicators of heightened cognitive load. In particular, we highlight the connection to the concepts of fluency and disfluency and the ways that previous research has proposed for discerning the ease with which language is produced.

In the data from our larger project, we anticipate that different segments of the narratives will demonstrate a variation in the presence of pauses and linguistic changes and that a cost-effective and less time-consuming way to capture this would be to divide the spoken and written processes into meaningful sequences and then explore the proportion of pauses, particularly those of extended duration, as well as the proportion of diverse and recurrent linguistic changes in each segment. Equally, the mere calculation of text length in each segment will give insights into the fluctuating nature of language production. Such an approach does not rule out subsequent or parallel qualitative analyses of the contexts of pauses and linguistic changes with a content-based focus, that is, the kind of annotations and analyses that require manual attention for accuracy and are thus more costly.

Further, in-depth examinations of pauses and changes during speaking and writing are time-consuming and require manual attention to ensure accuracy. From this point of view, it is imperative to explore methods that can be more cost-effective and, to a certain degree, automatized while still containing validity in annotations and categorizations of linguistic phenomena. While parts of spoken analyses may be automatically transcribed, filled pauses, repetitions, and reformulations will require manual annotation (even though speech recognition and AI have taken substantial leaps in the last decades). In turn, this also makes the analyses more time-consuming. In this regard, examining written accounts will have certain benefits; for example, the technical solution already exists for implementing a keylogger behind editor windows in report systems (Chukharev-Hudilainen et al. 2019) and both web-based and locally stored software is available (Wengelin and Johansson 2023). There is also existing software that can quickly provide an overview of the distribution of revision and pausing (in particular, see Inputlog, Leijten and Van Waes 2013).

Although specific solutions must be tailored for authorities and businesses that would want to use this possibility, there are fundamental technical solutions to build this on. In this way, less manual work is involved in collecting, annotating, and making initial analyses of

texts. Nevertheless, interpretation of data must be carried out manually to ensure validity. Also, the implementation of the method must be preceded by training in interpreting the findings. In this respect, the method of using data from real-time writing shares the same challenges as methods using spoken data. In addition, just as during the study of speaking, the study of real-time writing can be combined with methods for concurrently collecting auxiliary types of data, such as gaze behavior (cf. Johansson et al. 2023 for an overview of existing approaches).

4.5. Forensic Applications

Finally, what are the possible forensic applications of the present study? In our study, we have demonstrated that the examination of expressions of cognitive load during writing and speaking can serve as a mirror of instances, events, or circumstances that require extra thinking from the writer/speaker. With the assumption that deception induces increased cognitive load, such discoveries can be used for forensic purposes, although it is essential to first conduct more applied research to establish a baseline and variation across different populations and tasks. It is also important to note that this is one of many approaches that should be part of a forensic toolbox.

However, given that written reports are an essential part of many procedures in the legal system, such as requiring clients to give an initial report online in written form on a website, it is not unlikely that the examination of pausing and revision patterns during writing can be used as one of several indications of instances that need extra attention from interrogators and that it can, together with other evidence and circumstances, serve to inform about contexts where more evidence or investigations are necessary to collect. In the future, it may also be possible to ask witnesses to give written statements in a more secure environment (e.g., a police station) in an early stage of an investigation to capture less rehearsed accounts through keystroke logging.

To summarize, this article has shown and exemplified through one speaker and one writer that there are common phenomena associated with increased cognitive load in speaking and writing. Further, this article proposes methods for how such expressions can be investigated in a fruitful way across these modalities based on previous theories and results of research in the linguistic fields of real-time speaking and writing. The next step will be to investigate if there are any systematic differences between truthful and deceitful accounts in a larger pool of data, which is the goal of the larger project, which includes this study.

5. Conclusions

The foundational premise of this study posits that deceptive narrative accounts necessitate extensive planning, thereby impeding the fluency of the language production process and inducing increased cognitive load. Observable manifestations of this cognitive load in overt linguistic expression include disruptions, such as an augmented number of pauses, relatively longer pauses (filled and silent in spoken discourse), heightened revision through deleted characters and editing operations in writing, or an increased number of repetitions and rephrasing in speaking. While our primary objective was to establish a methodological framework for identifying signs of deceptive-induced cognitive load in language production, we were additionally focused on devising accessible and cost-effective approaches for this purpose.

In our examination of different approaches and methods to measure cognitive load in speaking as well as writing, we have consistently recognized that some methods necessitate more manual work than others. We advocate for the use of automatic or semi-automatic methods, whenever available, and highlight keystroke logging as a particularly valuable tool for investigating written language production. This preference is rooted in its ease of data collection, requiring comparatively little post-curation of data in contrast to transcription and annotation of spoken data. The keystroke logging software offers diverse output analysis files that facilitate the investigation of pauses and revisions.

Furthermore, exploring deception during writing provides a unique advantage. People are generally familiar with encountering final written texts but will normally not have observed the process of text composition. This lack of familiarity with writing processes makes it potentially easier to detect patterns of increased cognitive load associated with lying compared to spoken language production, where individuals may employ mimicking behaviors learned from countless spoken interactions.

While our methodological framework holds promise for practical forensic applications seen from a long-term perspective, such as employing keystroke logging tools to analyze written statements on web pages for witness reports, it is essential to acknowledge certain limitations. The larger study exclusively involves native language (L1) speakers, a deliberate choice made to maintain data integrity, as L2 speaking and writing often introduce heightened cognitive load. Additionally, the methodological approach assumes proficiency in keyboard-based writing, limiting its applicability. As an important note, the methodological framework we propose here is primarily applicable to computer-generated texts and leverages the analytical capabilities offered by current keystroke logging software programs (see Wengelin and Johansson 2023 for a comprehensive overview). While the foundational theories regarding how cognitive load impacts written language production are also relevant to handwriting (cf. van Hell et al. 2008), one should anticipate that differences in execution will come into play. For example, handwriting tends to be more time-consuming, and the process of revision is both more challenging and time-intensive, leaving more detectable traces. Expanding the framework's applicability to handwriting would necessitate additional research endeavors utilizing specialized tools designed for capturing and scrutinizing handwriting, such as Eye & Pen (Alamargot et al. 2006).

As discussed above, additional applications of the suggested methodology must be carried out before the method can be used in real-life contexts. Not the least, it is imperative to determine what role individual differences play and how one can establish a baseline for how cognitive load is expressed during truthful accounts. While the method of using keystroke logging eventually has the potential to be used in court, much more research is needed to establish various baselines concerning how deceptive behavior is manifested during writing in different contexts and for different individuals, where individual writing styles are important to consider. However, once such a body of research is established, this method can be used alongside other methods for information gathering.

In conclusion, we contend that writing can serve as a valuable complement to speaking as a forensic tool but cannot entirely replace it. Our methodological proposition seeks to expand the forensic toolbox. This endeavor holds significant forensic importance as individuals provide both truthful and deceptive narratives in both spoken and written formats. A thorough exploration of the processes associated with "speaking and writing truth and falsehood" and their interplay across modalities will offer valuable insights into forensic linguistics. Specifically, comparing the act of deception in spoken and written forms will illuminate best practices for extracting potentially deceptive information in various contexts, including witness accounts, security clearance interviews, and similar scenarios.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/languages9030085/s1. All data used in this study are available as Supplementary Materials.

Author Contributions: Conceptualization, K.G., V.J. and R.J.; methodology, K.G., V.J. and R.J; validation, K.G. and V.J.; formal analysis, K.G. and V.J.; investigation, K.G. and V.J.; resources, K.G. and V.J.; data curation, K.G.; writing—original draft preparation, K.G. and V.J.; writing—review and editing, K.G., V.J. and R.J.; visualization, K.G.; supervision, V.J. and R.J.; project administration, V.J.; funding acquisition, V.J. and R.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Marcus and Amalia Wallenberg Foundation, grant 2019.0058.

Institutional Review Board Statement: All methods were conducted in accordance with the Swedish Act concerning the Ethical Review of Research involving Humans (2003:460) and the Code of Ethics of the World Medical Association (Declaration of Helsinki). As established by Swedish authorities

and specified in the Swedish Act concerning the Ethical Review of Research involving Humans (2003:460), the present study does not require specific ethical review by the Swedish Ethical Review Authority due to the following reasons: (1) it does not deal with sensitive personal data, (2) it does not use methods that involve a physical intervention, (3) it does not use methods that pose a risk of mental or physical harm, (4) it does not study biological material taken from a living or dead human that can be traced back to that person.

Informed Consent Statement: Participants gave informed consent orally. These were recorded separately but in connection with the data collection. This consent form was chosen due to the pandemic situation, where we never met our participants in person. All informed consent included permission to publish in scientific journals, using data from the participants. Informed consent was obtained prior to the participation and all participants were informed that they could withdraw their participation at any time without any consequences. All participants are pseudonymized.

Data Availability Statement: The raw data used in this study are available in full as Supplementary Materials. These consist of final texts and corresponding linear files for the written material and transcriptions for the spoken material. The corresponding audio files can be provided upon request but will eventually be accessible in a corpus, available for research purposes.

Acknowledgments: In preparing elicitation material for this study, we used resources from Lund University Humanities Lab, and the authors want to gratefully acknowledge this support. We especially want to mention Peter Roslund for helping in producing the elicitation films, and Johan Frid for many discussions regarding decisions for data collection, transcription and analyses. We also want to acknowledge the important contribution from Sigrid Svensson who helped in data collection and transcription. We further acknowledge the participants in this study, who generously helped us by partaking. Finally, the idea to this project originated from exchanges and discusses within the think tank *Intelligent intelligence* at Lund University, and we are grateful for being connected to this environment. In the end we thank the Marcus and Amalia Wallenberg Foundation who through a generous grant made the whole project possible.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of this study, the collection, analyses, or interpretation of data, the writing of the manuscript, or the decision to publish the results.

References

Alamargot, Denis, David Chestnet, Cristophe Dansac, and Christine Ros. 2006. Eye and pen: A new device for studying reading during writing. *Behavior Research Methods, Instruments & Computers* 38: 287–99.

Alm, Per A. 2011. Cluttering: A neurological perspective. In *Cluttering: A Handbook of Research, Intervention and Education*. Edited by David Ward and Kathleen Scaler Scott. Psychology Press: pp. 3–28.

Alves, Rui A., and Teresa Limpo. 2015. Progress in Written Language Bursts, Pauses, Transcription, and Written Composition across Schooling. *Scientific Studies of Reading* 19: 374–91. [CrossRef]

Anthony, Laurence. 2023. *AntConc*. Tokyo: Waseda Univeristy.

Audacity. 2021. Audacity(R): Free Audio Editor and Recorder. (version 3.0.2). Available online: https://www.audacityteam.org/ (accessed on 24 February 2024).

Baddeley, Alan D. 2007. *Working Memory, Thought, and Action*. Oxford: Oxford University Press. [CrossRef]

Banerjee, Ritwik, Song Feng, Jun S. Kang, and Yejin Choi. 2014. Keystroke patterns as prosody in digital writing: A case study with deceptive reviews and essays. Paper presented at 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, October 25–29.

Barkaoui, Khaled. 2019. What can L2 Writers' Pausing Behavior Tell Us About Their L2 Writing Processes? *Studies in Second Language Acquisition* 41: 529–54. [CrossRef]

Barker, Megan S., Nicole L. Nelson, and Gail A. Robinson. 2020. Idea Formulation for Spoken Language Production: The Interface of Cognition and Language. *Journal of the International Neuropsychological Society* 26: 226–40. [CrossRef] [PubMed]

Berman, Ruth, and Ludo Verhoeven. 2002. Cross-linguistic perspectives on the development of text-production abilities: Speech and writing. *Written Language and Literacy* 5: 1–44. [CrossRef]

Biber, Douglas. 1988. *Variation Across Speech and Writing*. Cambridge, UK: Cambridge University Press.

Bott, Lewis, and Emma Williams. 2019. Psycholinguistic approaches to lying and deception. In *The Oxford Handbook of Lying*. Edited by Jörg Meibauer. Oxford: Oxford University Press, pp. 71–82.

Bourdin, Béatrice, and Michel Fayol. 1994. Is written language production more difficult than oral language production? A working memory approach. *International Journal of Psychology* 29: 591–620. [CrossRef]

Bowen, Niel. 2019. Unfolding choices in digital writing: A functional perspective on the language of academic revisions. *Journal of Writing Research* 10: 299–332. [CrossRef]
Brandt, Deborah. 2015. *The Rise of Writing: Redefining Mass Literacy*. Cambridge, UK: Cambridge University Press.
Buller, David B., and Judee K. Burgoon. 1996. Interpersonal Deception Theory. *Communication Theory* 6: 203–42. [CrossRef]
Burgoon, Judee K., and David B. Buller. 2008. Interpersonal Deception Theory. In *Engaging Theories in Interpersonal Communication: Multiple Perspective*. Edited by Leslie A. Baxter and Dawn O. Braithwaite. Thousand Oaks: Sage.
Chafe, Wallace, and Deborah Tannen. 1987. The Relation between Written and Spoken Language. *Annual Review of Anthropology* 16: 383–407. [CrossRef]
Chafe, Wallace, and Jane Danielwicz. 1987. *Properties of Spoken and Written Language. Technical Report No. 5*. Berkeley: California University, Berkeley Center for the Study of Writing.
Chafe, Wallace. 1994. *Discourse, Consciousness, and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing*. Chicago: University of Chicago Press.
Chanquoy, Lucile. 2009. Revision processes. In *The SAGE Handbook of Writing Development*. Edited by Roger Beard, Debra Myhill, Jeni Riley and Martin Nystrand. Thousand Oaks: SAGE, pp. 80–97.
Chenoweth, N. Ann, and John R. Hayes. 2001. Fluency in writing: Generating texts in L1. *Written Communication* 18: 80–98. [CrossRef]
Chenu, Florence, François Pellegrino, Harriet Jisa, and Michel Fayol. 2014. Interword and intraword pause threshold in writing. *Frontiers in Psychology* 5. [CrossRef] [PubMed]
Chukharev-Hudilainen, Evgeny, Aysel Saricaoglu, Mark Torrance, and Hui-Hsien Feng. 2019. Combined Deployable Keystroke Logging and Eyetracking for Investigating L2 Writing Fluency. *Studies in Second Language Acquisition* 41: 583–604. [CrossRef]
Clark, Herbert H. 1996. *Using Language*. Cambridge, UK: Cambridge University Press.
Clark, Herbert H., and Thomas Wasow. 1998. Repeating words in spontaneous speech. *Cognitive Psychology* 37: 201–42. [CrossRef] [PubMed]
Cleland, Alexandra A., and Martin J. Pickering. 2006. Do writing and speaking employ the same syntactic representations. *Journal of Memory and Language* 54: 185–98. [CrossRef]
Coleman, Linda, and Pal Kay. 1981. Prototype Semantics: The English Word Lie. *Language* 57: 26–44. [CrossRef]
Colwell, Kevin, Cheryl Hiscock-Anisman, Amina Memon, Alexis Rachel, and Lori Colwell. 2007. Vividness and spontaneity of statement detail characteristics as predictors of witness credibility. *American Journal of Forensic Psychology* 25: 5–30.
Cowan, Nelson. 2010. The Magical Mystery Four: How Is Working Memory Capacity Limited, and Why? *Current Directions in Psychological Science* 19: 51–57. [CrossRef]
Crystal, Thomas H., and Arthur S. House. 1990. Articulation rate and the duration of syllables and stress groups in connected speech. *The Journal of the Acoustical Society of America* 88: 101–12. [CrossRef]
Debey, Evelyne, Bruno Verschuere, and Geert Crombez. 2012. Lying and executive control: An experimental investigation using ego depletion and goal neglect. *Acta Psychologica* 140: 133–41. [CrossRef]
DePaulo, Bella M. 1992. Nonverbal behavior and self-presentation. *Psychological Bulletin* 111: 203–43. [CrossRef]
DePaulo, Bella M., James J. Lindsay, Brian E. Malone, Laura Muhlenbruck, Kelly Charlton, and Harris Cooper. 2003. Cues to deception. *Psychological Bulletin* 129: 74–118. [CrossRef] [PubMed]
Derrick, Douglas C., Thomas O. Meservy, Jeffrey L. Jenkins, Judee K. Burgoon, and Jay F. Nunamaker, Jr. 2013. Detecting Deceptive Chat-Based Communication Using Typing Behavior and Message Cues. *ACM Transactions on Management Information Systems (TMIS)* 4: 1–21. [CrossRef]
Duran, Nicholas D., Rick Dale, and Danielle S. McNamara. 2010. The action dynamics of overcoming the truth. *Psychonomic Bulletin & Review* 17: 486–91. [CrossRef]
Einarsson, Jan. 1978. *Talad Och Skriven Svenska: Sociolingvistiska Studier*. Lundastudier i Nordisk Språkvetenskap: Serie C, Studier i Tillämpad Nordisk Språkvetenskap 9; Lund: Ekstrand.
Eklund, Robert. 2004. Disfluency in Swedish Human–Human and Human-Machine Travel Booking Dialogues. Ph.D. dissertation, Linköping University, Linköping, Sweden.
Faigley, Lester, and Stephen Witte. 1981. Analyzing Revision. *College Composition and Communication* 32: 400–14. [CrossRef]
Feng, Ruiling, and Qian Guo. 2022. Second language speech fluency: What is in the picture and what is missing. *Frontiers in Psychology* 13: 859213. [CrossRef] [PubMed]
Fitzgerald, Jill. 1987. Research on revision in writing. *Review of Educational Research* 57: 481–506. [CrossRef]
Flower, Linda S., and John R. Hayes. 1981. A Cognitive Process Theory of Writing. *College Composition and Communication* 32: 365–87. [CrossRef]
Fox Tree, Jean E. 2001. Listeners' Uses of Um and Uh in Speech Comprehension. *Memory & Cognition* 29: 320–26. [CrossRef]
Fox Tree, Jean E., and Herbert H. Clark. 1997. Pronouncing the as thee to signal problems in speaking. *Cognition* 62: 151–67. [CrossRef] [PubMed]
Fox Tree, Jean E., and Josef C. Schrock. 2002. Basic Meanings of You Know and I Mean. *Journal of Pragmatics: An Interdisciplinary Journal of Language Studies* 34: 727–47. [CrossRef]
Fromkin, Victoria S., ed. 1973. *Speech Errors as Linguistic Evidence, Janua Linguarum. Series Maior: 77*. The Hauge: Mouton.
Gärdenfors, Moa, and Victoria Johansson. 2023. Written products and writing processes in Swedish deaf and hard of hearing children: An explorative study on the impact of linguistic background. *Frontiers in Psychology* 14: 1112263. [CrossRef] [PubMed]

Goldman Eisler, Frieda. 1968. Psycholinguistics: Experiments in Spontaneous Speech. *Lingua: International Review of General Linguistics* 25: 152–64.
Goupil, Louise, Emmanuel Ponsot, Daniel Richardson, Gabrial Reyes, and Jean-Julien Aucouturier. 2021. Listeners' perceptions of the certainty and honesty of a speaker are associated with a common prosodic signature. *Nature Communications* 12: 861. [CrossRef] [PubMed]
Grabowski, Joachim. 2008. The internal structure of university students' keyboard skills. *Journal of Writing Research* 1: 27–52. [CrossRef]
Granhag, Pär Anders, Aldert Vrij, and Bruno Verschuere. 2015. *Detecting Deception: Current Challenges and Cognitive Approaches*. Wiley Series in Psychology of Crime, Policing and Law; Chichester: Wiley-Blackwell.
Halliday, Michael A. K. 1985. *Spoken and Written Language*. Geelong: Deakin University Press.
Hayes, John R., and N. Ann Chenoweth. 2006. Is working memory involved in the transcribing and editing of texts? *Written Communication* 23: 135–49. [CrossRef]
Hayes, John R., Linda Flower, Karen A. Schriver, James Stratman, and Linda Carey. 1987. Cognitive processes in revision. In *Advances in Applied Psycholinguistics: Vol 2 Reading, Writing, and Language Processing*. Edited by Sheldon Rosenberg. Cambridge, UK: Cambridge University Press, pp. 176–241.
Heldner, Mattias, and Jens Edlund. 2010. Pauses, gaps and overlaps in conversations. *Journal of Phonetics* 38: 555–68. [CrossRef]
Hunt, Kellogg W. 1970. Syntactic maturity in schoolchildren and adults. *Monographs of the Society for Research in Child Development* 35: 1–67. [CrossRef]
Johansson, Roger, Åsa Wengelin, Victoria Johansson, and Kenneth Holmqvist. 2010. Looking at the keyboard or the monitor: Relationship with text production processes. *Reading and Writing* 23: 835–51. [CrossRef]
Johansson, Victoria, Merle Horne, and Sven Strömqvist. 2001. *Final Aspiration as a Phrase Boundary Cue in Swedish: The Case of Att 'That'*. Working Papers. Lund: Lund University, Department of Linguistics and Phonetics.
Johansson, Victoria, Roger Johansson, and Åsa Wengelin. 2023. Using keystroke logging for studying L2 writing processes. In *Research Methods in the Study of Writing Processes*. Edited by Rosa M. Manchón and Julio Roca de Larios. Amsterdam: John Benjamins, pp. 161–82.
Johansson, Victoria. 2000. Developing Editing. In *Developing Literacy across Genres, Modalities and Languages*. Edited by Melina Aparici, Noemi Argerich, Joan Perera, Elisa Rosadom and Liliana Tolchinsky. Barcelona: University of Barcelona.
Johansson, Victoria. 2009. Developmental Aspects of Text Production in Writing and Speech. Ph.D. thesis, Department of Linguistics and Phonetics, Centre for Languages and Literature, Lund University, Lund, Sweden.
Jong, Nivja H. 2016. Fluency in second language assessment. In *Handbook of Second Language Assessment*. Edited by Dina Tsagari and Jayanti Banerjee. Boston: De Gruyter Mouton, pp. 203–18.
Kaufer, David S., John R. Hayes, and Linda Flower. 1986. Composing Written Sentences. *Research in the Teaching of English* 20: 121–40.
Kellogg, Ronald, Casey E. Turner, Alison P. Whiteford, and Andrew Mertens. 2016. The role of working memory in planning and generating written sentences. *Journal of Writing Research* 7: 397–416. [CrossRef]
Kellogg, Ronald. 2008. Training writing skills: A cognitive developmental perspective. *Journal of Writing Research* 1: 1–26. [CrossRef]
Knapp, Mark L., Roderick P. Hart, and Harry S. Dennis. 1974. An exploration of deception as a communication construct. *Human Communication Research* 1: 15–29. [CrossRef]
Leijten, Marielle, and Luuk Van Waes. 2013. Keystroke Logging in Writing Research: Using Inputlog to Analyze and Visualize Writing Processes. *Written Communication* 30: 358–92. [CrossRef]
Leins, Drew A., Ronald P. Fisher, and Aldert Vrij. 2012. Drawing on Liars' Lack of Cognitive Flexibility: Detecting Deception Through Varying Report Modes. *Applied Cognitive Psychology* 26: 601–7. [CrossRef]
Levelt, Willem J. M. 1989. *Speaking: From Intention to Articulation*. Cambridge, UK: MIT Press.
Lindgren, Eva, and Kirk P. H. Sullivan, eds. 2019. *Observing Writing: Insights from Keystroke Logging and Handwriting, Studies in Writing: Volume 38*. Leiden and Boston: Brill.
Lindgren, Eva, Asbjørg Westum, Hanna Outakoski, and Kirk P. H. Sullivan. 2019. Revising at the Leading Edge: Shaping Ideas or Clearing up Noise. In *Observing Writing: Insights from Keystroke Logging and Handwriting*. Edited by Eva Lindgren and Kirk P. H. Sullivan. Leiden: Brill, pp. 346–65.
Lindgren, Eva. 2005. Writing and Revising: Didactic and Methodological Implications of Keystroke Logging. Ph.D. thesis, Umeå University, Umeå, Sweden.
Linell, Per. 2009. *Rethinking Language, Mind, and World Dialogically: Interactional and Contextual Theories of Human Sense-making. Advances in Cultural Psychology*. Charlotte: Information Age Publication.
Lively, Scott E., David B. Pisoni, W. Van Summers, and Robert H. Bernacki. 1993. Effects of cognitive workload on speech production: Acoustic analyses and perceptual consequences. *The Journal of the Acoustical Society of America* 93: 2962–73. [CrossRef] [PubMed]
Loban, Walter. 1976. *Language Development: Kindergarten through Grade Twelve. Re-Search Report No. 18*. Urbana: National Council of Teachers of English.
MacWhinney, Brian. 2000. *The CHILDES Project: Tools for Analyzing Talk*, 3rd ed. Mahwah: Lawrence Erlbaum Associates.
Manchón, Rosa M. 2020. *Writing and Language Learning: Advancing Research Agendas*. Amsterdam: John Benjamins.
Manchón, Rosa M., and Julio Roca de Larios, eds. 2023. *Research Methods in the Study of Writing Processes, Research Methods in Applied Linguistics (RMAL)*. Amsterdam: John Benjamins.

Mann, Samantha. 2019. Lying and Lie Detection. In *The Oxford Handbook of Lying*. Edited by Jörg Meibauer. Oxford: Oxford University Press, pp. 408–19.
Matsuhashi, Ann. 1981. Pausing and Planning: The Tempo of Written Discourse Production. *Research in the Teaching of English* 15: 113–34.
McCutchen, Deborah. 1996. A Capacity Theory of Writing: Working Memory in Composition. *Educational Psychology Review* 8: 299–325. [CrossRef]
McCutchen, Deborah. 2000. Knowledge, Processing, and Working Memory: Implications for a Theory of Writing. *Educational Psychologist* 35: 13–23. [CrossRef]
Murray, Donald M. 1978. Internal revision: A process of discovery. In *Research on Composing: Points of Departure*. Edited by Charles R. Cooper and Lee Odell. Urbana: National Council of Teachers of English.
Newman, Matthew L., James W. Pennebaker, Diane S. Berry, and Jane M. Richards. 2003. Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin* 29: 665–75. [CrossRef]
Nishimura, Fumiko. 2018. Lying in Different Cultures. In *The Oxford Handbook of Lying*. Edited by Jörg Meibauer. Oxford: Oxford University Press.
Norrby, Catrin. 2014. *Samtalsanalys: Så gör vi när vi Pratar med Varandra*, 3 [rev.] ed. Lund: Studentlitteratur.
Nottbusch, Guido. 2010. Grammatical planning, execution and control in written sentence production. *Reading and Writing* 23: 777–801. [CrossRef]
Ofen, Noa, Susan Whitfield-Gabrieli, Xiaoqian J. Chai, Rebecca F. Schwarzlose, and John D. E. Gabrieli. 2017. Neural correlates of deception: Lying about past events and personal beliefs. *Social Cognitive & Affective Neuroscience* 12: 116–27. [CrossRef]
Olive, Thierry. 2014. Toward a parallel and cascading model of the writing system: A review of research on writing processes coordination. *Journal of Writing Research* 6: 173–94. [CrossRef]
Ortega, Lourdes. 2009. *Understanding Second Language Acquisition*. Understanding Language Series; London: Hodder Education.
Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Association for Computational Linguistics (ACL) System Demonstrations*. Dublin: Association for Computational Linguistics.
Repke, Meredith A., Lucian Gideon Conway, and Shannon C. Houck. 2018. The Strategic Manipulation of Linguistic Complexity: A Test of Two Models of Lying. *Journal of Language and Social Psychology* 37: 74–92. [CrossRef]
Sacks, Harvey, Emanuel A. Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language* 50: 696–735. [CrossRef]
Schreiber, Kayleen E., and Bob McMurray. 2019. Listeners can anticipate future segments before they identify the current one. *Attention, Perception, & Psycholinguistics* 81: 1147–66. [CrossRef]
Scott, Cheryl M. 1988. Spoken and written syntax. In *Later Language Development. Ages Nine through Nineteeen*. Edited by Marilyn A. Nippold. Austin: Elsevier, pp. 49–95.
Seymour, Travis L. 2001. A EPIC Model of the 'Quilty Knowledge Effect': Strategic and Automatic Processes in recognition. Ph.D. thesis, The University of Michigan, Ann Arbor, MI, USA.
Sokolov, Eugene N. 1963. *Perception and the Conditioned Reflex*. New York: Macmillan.
Song, Shuxian, and Dechao Li. 2020. The predicting power of cognitive fluency for the development of utterance fluency in simultaneous interpreting. *Frontiers in Psychology* 11: 1864. [CrossRef] [PubMed]
Sörqvist, Patrik, Anatole Nöstil, and Niklas Halin. 2012. Disruption of writing processes by the semanticity of background speech. *Scandinavian Journal of Psychology* 53: 97–102. [CrossRef] [PubMed]
Spelman Miller, Kristyan. 2006a. Pausing, productivity and the processing of topic in online writing. In *Computer Keystroke Logging and Writing: Methods and Applications*. Edited by Kirk P. H. Sullivan and Eva Lindgren. Amsterdam: Elsevier, pp. 131–56.
Spelman Miller, Kristyan. 2006b. The pausological study of written language production. In *Studies in Writing, Vol 18, Computer Keystroke Logging: Methods and Apllications*. Edited by Kirk P. H. Sullivan and Eva Lindgren. Oxford: Elsevier, pp. 11–30.
Sporer, Siegfried Ludwig, and Barbara Schwandt. 2006. Paraverbal Indicators of Deception: A Meta-analytic Synthesis. *Applied Cognitive Psychology* 20: 421–46. [CrossRef]
Sporer, Siegfried Ludwig, and Barbara Schwandt. 2007. Moderators of nonverbal indicators of deception: A meta-analytic synthesis. *Psychology, Public Policy, and Law* 13: 1–34. [CrossRef]
Stevenson, Marie, Rob Schoonen, and Kees de Glopper. 2006. Revising in two languages: A multi-dimensional comparison of online writing revisions in L1 and FL. *Journal of Second Language Writing* 15: 201–33. [CrossRef]
Suchotzki, Kristina, Bruno Verschuere, Bram Van Bockstaele, Gershon Ben-Shakhar, and Geert Crombez. 2017. Lying takes time: A meta-analysis on reaction time measures of deception. *Psychological Bulletin* 143: 428–53. [CrossRef]
Svenska Akademiens Ordlista över svenska språket, SAOL 14. 2015. Stockholm: Svenska akademien.
Talwar, Victoria. 2019. Development of Lying and Cognitive Abilities. In *The Oxford Handbook of Lying*. Edited by Jörg Meibauer. Oxford: Oxford University Press, pp. 399–407.
Torrance, Mark, Roger Johansson, Victoria Johansson, and Åsa Wengelin. 2016. Reading during the composition of multi-sentence texts: And eye-movement study. *Psychological Reserch* 80: 729–43. [CrossRef]
Torrance, Mark. 2016. Understanding planning in text production. In *Handbook of Writing Research*. Edited by Charles A. MacArthur, Steve Graham and Jill Fitzgerald. New York: Guilford.

Undeutsch, Udo. 1989. The development of statement reality analysis. In *Credibility Assessment*. Edited by J. Yullie. Dordrech: Kluwer Academic Publishers, pp. 101–19.
van Hell, Janet G., Ludo Verhoeven, and Liesbeth M. van Beijsterveldt. 2008. Pause time patterns in writing narrative and expository texts by children and adults. *Discourse Processes* 45: 406–27. [CrossRef]
Van Waes, Luuk, and Mariëlle Leijten. 2015. Fluency in Writing: A Multidimensional Perspective on Writing Fluency Applied to L1 and L2. *Computers and Composition* 38: 79–95. [CrossRef]
Van Waes, Luuk, Marielle Leijten, Jens Roeser, Tierry Olive, and Joachim Grabowski. 2021. Measuring and assessing typing skills in writing research. *Journal of Writing Research* 13: 107–53. [CrossRef]
Vrij, Aldert, Katherine Edward, Kim P. Roberts, and Ray Bull. 2000. Detecting Deceit via Analysis of Verbal and Nonverbal Behavior. *Journal of Nonverbal Behavior* 24: 239–63. [CrossRef]
Vrij, Aldert, Pär Anders Granhag, Tzachi Ashkenazi, Giorgio Ganis, Sharon Leal, and Ronald P. Fisher. 2022. Verbal Lie Detection: Its Past, Present and Future. *Brain Sciences* 12: 1644. [CrossRef] [PubMed]
Vrij, Aldert, Samantha A. Mann, Ronald P. Fisher, Sharon Leal, Rebecca Milne, and Ray Bull. 2008. Increasing Cognitive Load to Facilitate Lie Detection: The Benefit of Recalling an Event in Reverse Order. *Law and Human Behavior* 32: 253–65. [CrossRef] [PubMed]
Vrij, Aldert, Samantha Mann, Sharon Leal, and Ronald Fisher. 2010. 'Look into my eyes': Can an instruction to maintain eye contact facilitate lie detection? *Psychology, Crime and Law* 16: 327–48. [CrossRef]
Walczyk, Jeffrey J., Frank D. Igou, Lexie P. Dixon, and Talar Tcholakian. 2013. Advancing lie detection by inducing cognitive load on liars: A review of relevant theories and techniques guided by lessons from polygraph-based approaches. *Frontiers in Psychology* 4: 14. [CrossRef]
Walczyk, Jeffrey J., Jonathan P. Schwartz, Rayna Clifton, Barett Adams, Min Wei, and Peijia Zha. 2005. Lying person to person about life events: A cognitive framework for lie detection. *Personnel Psychology* 58: 141–70. [CrossRef]
Walczyk, Jeffrey J., Karen S Roper, Eric Seemann, and Angela M. Humphrey. 2003. Cognitive mechanisms underlying lying to questions: Response time as a cue to deception. *Applied Cognitive Psychology* 17: 755–74. [CrossRef]
Walczyk, Jeffrey J., Kevin T. Mahoney, Debbis Doverspike, and Diana A. Griffith-Ross. 2009. Cognitive lie detection: Response time and consistency of answers as cues to deception. *Journal of Business and Psychology* 24: 33–49. [CrossRef]
Wengelin, Åsa, and Victoria Johansson. 2023. Investigating writing processes with keystroke logging. In *Digital Writing Technologies: Impact on Theory, Research, and Practice in Higher Education*. Edited by Otto Kruse, Christian Rapp, Chris M. Anson, Kalliopi Benetos, Elena Cotos, Ann Devitt and Antonnete Shibani. Berlin: Springer.
Wengelin, Åsa, Roger Johansson, Johan Frid, and Victoria Johansson. 2023. Capturing writers' typing while visually attending the emerging text: A methodological approach. *Reading and Writing: An Interdisciplinary Journal* 37: 265–89. [CrossRef]
Wengelin, Åsa. 2006. Examining pauses in writing: Theory, methods and empirical data. In *Computer Key-Stroke Logging: Methods and Applications*. Edited by Kirk P. H. Sullivan and Eva Lindgren. Amsterdam: Elsevier, pp. 107–30.
Williams, Emma J., Lewis A. Bott, John Patrick, and Michael B. Lewis. 2013. Telling Lies: The Irrepressible Truth? *PLoS ONE* 8: e60713. [CrossRef] [PubMed]
Zuckerman, Miron, Bella M. DePaulo, and Robert Rosenthal. 1981. Verbal and Nonverbal Communication of Deception. *Advances in Experimental Social Psychology* 14: 1–59.

languages

MDPI

Article

I, as a Fault—Condemnation of Being and Power Dynamics in the Parent-Child Interaction †

Rose Moreau Raguenes

AGORA Laboratory EA 7392, Institute for Digital Humanities FED 4284, CY Cergy Paris Université, 95000 Cergy, France; rose.moreau-raguenes@cyu.fr

† This article is a revised and expanded translation of Moreau Raguenes, Rose. L'être comme fautif. Actes de condamnation, altérité et rapport de places dans l'interaction parent-enfant. *SHS Web of Conferences*, 191, article 01009. 2024. https://doi.org/10.1051/shsconf/202419101009.

Abstract: This article explores the power dynamics underlying verbal abuse within the parent-child interaction. Through a reception-based approach, it focuses on condemnation acts of being (e.g., *you are a good for nothing*) directed by abusive parents towards their children and reported by the latter in anonymous testimonies published on the Francophone Instagram account *Parents toxiques*; a sample of ten testimonies is examined. The analyses conducted show that (i) the ontological assertion of power over the other is constructed from the predicative level, with processes that concern their being in its entirety and present condemnation as an objective reality. (ii) The condemnation of being draws its pragmatic force from its legitimisation—by relying on norms presented as self-evident and universal and by highlighting the harm caused by the other. (iii) As a speaker, constructing the other's being as at fault involves, to varying degrees, essentialising and downgrading them as well as conflating their intrinsic worth with one's beliefs and needs. In conclusion, the notion of condemnation acts of being—along with its descriptors—provides an effective framework that can be applied to reports and direct observations to help various professionals identify and assess transgressions and/or dysfunctions in authority relationships.

Keywords: verbal abuse; child abuse; testimony; discourse analysis; argumentation; social media; speech acts; politeness; impoliteness; interaction

Academic Editor: Zhengrui Han

Received: 14 July 2024
Revised: 4 March 2025
Accepted: 7 March 2025
Published: 19 March 2025

Citation: Moreau Raguenes, R. (2025). I, as a Fault—Condemnation of Being and Power Dynamics in the Parent-Child Interaction. *Languages*, *10*(3), 54. https://doi.org/10.3390/languages10030054

1. Introduction

1.1. Approaching Child Abuse

In psychology, experimental studies have been conducted—notably by John and Julie Gottman and their colleagues—to identify predictors of relationship success or failure among couples (see J. Gottman & Gottman, 2017 for a summary). Four "destructive relationship behaviours" were identified: criticism, defensiveness, contempt, and stonewalling (J. M. Gottman et al., 2019). Although our purpose is not to characterise what a successful parent-child relationship might be or to predict its longevity, the present study resonates with these findings. It aims to deepen the understanding of verbal abuse within parent-child interactions by studying, through discourse and interaction, how it may be committed and perceived.

The research objective presents several challenges. I will not insist on the challenges of conducting research involving children, particularly for ethical and deontological reasons, which have been widely documented across various fields (e.g., Kopelman, 2000; Einarsdóttir, 2007; Kousholt & Juhl, 2023). In linguistics, studying child abuse empirically presents specific challenges. The first category of difficulties relates to accessing authentic

parent-child interactions, especially when it comes to collecting informed consent and addressing the "observer's paradox" (Labov, 1972, p. 209). Researchers have tried to overcome this by giving participants microphones to record authentic discourse within the family sphere (e.g., Laforest, 2002 on complaining in everyday conversation; Clancy, 2011, on hedging in family discourse). However, in studies focusing on abusive interactions, obtaining informed consent could introduce biases, and the anonymity of participants might be compromised if harmful behaviour is observed. Another major difficulty when researching violence is "knowing whether the researcher should consider as violent acts that the participants do not identify as such, and vice-versa"[1], as Ayimpam (2015, §2) points out. In other words, from which viewpoint should parent-child interactions be categorised as abusive? The instability and situatedness of the category *abuse* affects the entire analytical process—from constructing a corpus to delineating what could and could not be analysed in that corpus.

Given the ethical and methodological challenges of accessing authentic abusive parent-child interactions, this study proposes an alternative approach. Primarily rooted in the French context, it is part of a research project that studies the sociodiscursive representation of child abuse in anonymous testimonies published on a Francophone Instagram account, *Parents toxiques* ('Toxic Parents')[2]. Since abuse is an unstable, context-sensitive category, clarifying the current French legislation is necessary: as redefined by the 2019 law on Ordinary Educational Violence, parental authority "shall be exercised without physical or psychological violence" (Article 371-1 of the *Code civil*)[3]. This law was, notably, not associated with new sanctions. It was meant to prevent judges from invoking a jurisprudentially established "right of correction" on the part of parents and, more generally, to officially affirm that violence could no longer be considered an educational tool. Moreover, the fact that psychological violence is considered part of violence and is punishable by law (Articles 222-14-3 and 222-33-2-2 of the *Code Pénal*)[4] raises the issue of its objectivisation for professionals and individuals who may experience it.

In light of this recent legal redefinition of parental authority in France, the present study engages with forensic linguistics on two levels: (i) it uses linguistic descriptors to characterise verbal behaviours that are now delegitimised and/or prohibited by French law; and (ii) it does so by focusing on their publicisation, i.e., how certain parental behaviours are publicly shared and represented on social media—which echoes the origins of *forensic*, 'pertaining to the forum'. To this end, I shall introduce key notions derived from studies on verbal violence, conducted by francophone analysts in interactional Sociolinguistics and Discourse analysis (namely Laforest & Vincent, 2004; Rosier, 2009; Vincent, 2013; Laforest & Moïse, 2013; Moïse et al., 2019).

1.2. Verbal Violence Framework and Objectives of the Study

Fracchiolla et al. (2023, §3) define verbal violence as "a transgression into the other person's territory, through speech, against their consent [. . .]" which "constructs unequal relationships and relies on domination, oppositions, and power dynamics". This is notably carried out through condemnation acts[5] (Laforest & Moïse, 2013), i.e., "forms of verbal violence that undermine the identity of others" (Fracchiolla et al., 2023, §16).

> "Speech acts of condemnation, such as provocation, threats, reproaches, or insults, are at the core of perceived effects of verbal violence because they aim to affect the other person, to alter their sense of security, dignity, and/or social esteem, and to demean them by asserting power through pragmatic means. These acts are often accompanied by argumentative devices that legitimise the judgements issued ("I treat you this way because. . .")". (Fracchiolla et al., 2023, §13)

The study of condemnation acts is rooted in a conception of language as action (Austin, 1962; Searle, 1982), and a (socio)linguistic perspective provides analytical tools to "objectively address what occurs in violent discourse" (Fracchiolla et al., 2023, §5). To account for the dynamics of verbal violence, Laforest and Moïse (2013) conceptualise a shift from a conflict over an object to a conflict over persons, and from the condemnation of doing to the condemnation of being (Figure 1):

Conflict over an object	**Conflict over persons**		
	condemnation of doing →		**condemnation of being**
There's too much noise	*You're making too much noise*	*You're being too noisy (right now)*	*You're noisy (in general)*

Figure 1. Types of Conflicts and Condemnation Acts (Laforest & Moïse, 2013, p. 90, translated).

The authors define condemnation acts of doing as "speech acts through which a speaker expresses dissatisfaction with an act or behaviour of an individual that they deem to be inadequate"; when condemnation shifts to focus on the being, one "detaches the contentious behaviour from its particular space and time" and "essentialises the condemnation", making it a "permanent characteristic of the condemned individual" (Laforest & Moïse, 2013, p. 89). This shift was also identified as part of the "destructive relationship behaviours" mentioned earlier (J. M. Gottman et al., 2019, p. 1212):

> "Criticism happens when someone verbally attacks their partner, placing the blame for whatever problem they are experiencing inside the partner's character. Instead of complaining about the situation and offering a way to make things better, the user of criticism communicates a belief that **the problem is occurring because of a defect in their partner**. Words such as "always" and "never" frequently appear in criticism-based statements".

Since this shift is a central feature in verbally violent or dysfunctional interactions, this article proposes to delve deeper into its dynamics: focusing on the condemnation of being should allow us to deepen our understanding of its functioning. First, its linguistic and discursive materiality warrants a thorough examination to characterise it at different levels and identify key descriptors and devices. Second, the condemnation of being is represented by Laforest and Moïse (2013) as the extreme end of a continuum; there may be different degrees or variations of intensity within the condemnation of being. Finally, effects are not easily captured in discourse analysis, as studying the reception of a given discourse requires additional data; analysing spontaneous online testimonies can give us access to the long-term reception of verbal violence. With this in mind, we shall analyse the condemnation acts of being addressed by the abusive parent to their child, as reported by the latter in the testimonies, to describe the underlying treatment of otherness.

This study of the condemnation of being cannot be separated from a reflection on the interactional dynamics inherent in parent-child interactions. While every interaction involves shifts in terms of dynamics, the parent-child interaction is characterised by an inherent asymmetry "by virtue of the very status of the participants" (Laforest & Vincent, 2006, p. 8)—with different ages and stages of development, material dependency, etc.[6] The interactional positions are complementary, with the parent assuming the role of the guarantor and the child having a status that ensures protection and guidance. These interactional dynamics correspond to an authority relationship, that is, a "relationship of mutual recognition in an asymmetrical situation" where one participant is in a higher position and the other in a lower position (Moïse et al., 2019, pp. 25–26). Such a relationship

"is not to be confused with a relationship of domination as it is a co-constructed and reassuring relationship" (Moïse et al., 2019, p. 26).

Taking into consideration these interactional dynamics, the starting point of our reflection will be the following: is the authority relationship disrupted in and by the condemnation acts of being, and if so, how? The discursive analysis of the acts of condemnation reported in the testimonies will examine the ways in which the treatment of otherness contributes to a perception of the interaction as inappropriate or abusive. To this end, I present the methodological and analytical framework of the study in Section 2; this section will seek to clarify how to identify and analyse the reported condemnation acts. Section 3 will then highlight the discursive mechanisms of the condemnation of being reported by the authors of the testimonies. Finally, Section 4 will focus on identifying the reasons underlying the construction of one's being as at fault and their implications in terms of interactional dynamics.

2. Studying Reported Speech Acts in Testimonies: Methodology of Collection and Analysis

The *Parents toxiques* account, created in July 2019, is public and features content related to child abuse, discrimination, and resilience more broadly. Testimonies are submitted to the account creator via direct message; they are then published without revealing the author's identity and techno-contextualised (Longhi, 2013) with the hashtags #*temoignage* #*parentstoxiques*. In total, 350 anonymous testimonies were published from July 2019 to September 2021, that is, from the account creation to a temporary interruption of its activity.

2.1. Sampling Methodology

This study focuses on a sample of ten anonymous testimonies published between May 2020 and July 2021, written by nine women and one man. Since the testimonies are anonymous, the authors' gender was inferred from grammatical and lexical gender markers (e.g., agreements, common nouns such as *daughter*). The first names used to refer to the authors are pseudonyms assigned for this project[7]. Starting from the most recently published testimonies, the criteria presented in Figure 2 were applied to create a sample of ten texts: the shortest testimony comprises 414 words, and the longest 996 words.

At this stage of the research project, these criteria were not intended to ensure a representative sample of testimonies—which could have been achieved by randomly selecting ten testimonies. Rather, the aim was to constitute a sample that would enable us to examine retrospective recountings of abusive behaviours, experienced in relation to the authors' biological or adoptive parents; the sampling criteria were therefore defined based on the project's objectives. The sampling process provided insight into both the diversity and specific features of the *Parents toxiques* testimonies—for instance, a large majority of the testimonies are written by women, according to grammatical and lexical markers.

To study the speech acts of abusive parents in these testimonies, two methodological and analytical issues should now be addressed: delimiting reported speech acts, and categorising them as condemnation acts.

Format
1. The testimony must be at least fifty lines long (in the original layout). This minimum length criterion helps isolate testimonies that engage in storytelling and develop the events being recounted; it aligns with criterion 5.
Content
2. The testimony focuses on abuse committed against the author. Selecting testimonies that focus on violence against others would lead to analysing narratives of abuse with different implications. Therefore, testimonies in which the author was not a direct victim of the parent – i.e. those that *only* described violent behaviours toward a third party – were excluded.
3. The abuse was committed by the biological or adoptive parents. This study explores the retrospective representation of an abusive relationship featuring a parent-child bond.
4. The abuse took place during childhood and/or adolescence (up to eighteen years old). This criterion is based on French law – parental abuse concerns minors. Testimonies that describe abnormal parental behaviors in adulthood were excluded unless they also depicted abusive behaviours during childhood and/or adolescence (based on the temporal information provided by the author).
5. The testimony is a narrative and recounts a specific story from the time of abuse. It was important to observe and characterise how the author represents themselves within their story.
6. The author does not seek legal or other forms of assistance from the community within the testimony. Explicitly requesting information or help would assign a different purpose to the act of testifying.
7. The testimony includes reported speech from the parent. This was essential to study how the parent's speech was represented, as one of the research focuses is on speech acts perceived as violent.
8. The testimony does not concern sexual violence. Addressing sexual violence would have required a distinct theoretical and analytical framework.
Author characteristics
9. The author must have left the abusive situation. As this study focuses on the meaning that speakers attribute to their experience of abuse in retrospect, the person cannot still be in the abusive situation at the time of writing; testimonies where the author had just left the abusive situation were also excluded.

Figure 2. Sampling Criteria (Adapted from Moreau Raguenes, 2022).

2.2. Delimitating the Parent's Reported Speech

First, we need to define and delineate the forms that may fall under the reported speech acts. I rely on Rosier's definition of reported speech as "ways of reporting, representing, interpreting, and circulating someone else's discourse by configuring a relationship between a discourse that creates a particular enunciative space (the citing discourse) and a discourse that is set apart and—univocally or not—attributed to another source (the cited discourse)" (Rosier, 2008, p. 137). More specifically, our focus is on the reported speech of the abusive parent(s) in which a speech act is clearly addressed to the speaker (i.e., the author of the testimony). The following forms are therefore excluded:

- Acts of insult, denigration, reproach, etc., reported without "the linguistic context containing any semes directly pertaining to the speech act" (Rosier, 1999, p. 129).
  [1] Whether it was **humiliation** in public places, **insults, denigration**, psychological pressure or repeated crises, coming to his house became an ordeal.
  *Qu'il s'agisse d'**humiliations** dans des lieux publiques, **insultes, dénigrements**, pressions psychologiques et crises à répétition, venir chez lui devenait un calvaire. (Albane)*[8]
- Cases in which identifying the enunciative source and/or interpreting the segment as reported speech present challenges[9]:
  [2] In primary school, my father hit me when I **"made him angry"**
  *En primaire, mon père me frappait lorsque je **"le mettais en colère"** (Chiara)*
  [3] My brother had serious learning and behavioural disorders. So for me, **it was better not to "make things worse"**. Not to make waves. Not to be an extra burden.

> *Mon frère avait de gros troubles de l'apprentissage et du comportement. Alors moi,* ***il valait mieux que j'en "rajoute pas".*** *Que je ne fasse pas de vagues. Que je ne sois pas un fardeau supplémentaire. (Théa).*

The age of the speakers at the time they received these speech acts, whether as children (up to eighteen years in France) or adults, also warrants careful consideration. Applying an age limit to select and exclude speech acts for analysis raises two methodological challenges. Firstly, a lack of available information: the testimonies do not clearly indicate the age of the authors when they received a given act. Furthermore, the speech acts reported by the speakers have often been performed more than once, potentially at different ages. Beyond the lack of access to the situation of utterance, excluding speech acts produced in adulthood would restrict the analysis to the legal distinction between minors and adults. The authors of the testimonies frequently address the ongoing parent-child relationship into adulthood, however. This is one of the specificities of this corpus, and the representation of condemnation acts directed at the adult child can also provide insight into what constitutes an inappropriate power dynamic between parent and child according to the speakers. Therefore, reported speech acts that appeared to have been produced when the speaker had reached adulthood were included.

2.3. Categorising Reported Speech Acts as Condemnation Acts

Since we do not have access to the primary interaction in these testimonies, but only to the parents' reported speech, we need to determine how to identify speech acts as condemnation acts—in other words, from which perspective it is possible to categorise them as verbally violent.

Vincent (2013, p. 38) observes that "each device that causes a disruption in expectations has two poles—one acceptable, even recommended or salutary, and the other unacceptable: ritual insult and personal insult, warning and threat, critique and denigration, as well as all intermediate interpretations". As analysts, we must therefore dissociate "the linguistic means that can potentially cause violence from violence itself" (Vincent et al., 2008, cited by Vincent, 2013, p. 38), and be careful not to judge for ourselves what is violent or acceptable. Apart from the analyst's judgement, another possible approach would be to consider the viewpoint and intention of the speakers—in this case, the parents. However, as Moïse et al. (2019, p. 132) point out, speakers do not necessarily have the intention or even the awareness of acting as a conduit for verbal violence, and we cannot access the speakers' intentions in discourse analysis and sociolinguistics: "The subject who exhibits extreme verbal aggression is primarily driven by impulses that overwhelm them. Consequently, detecting the speakers' possible intentions remains challenging as neither the addressee nor the sociolinguist has direct access to the subjects' intentions". The present work thus follows Laforest and Moïse (2013, p. 91) in their analytical choice to base the analysis on the viewpoint of reception:

> "**[. . .] the viewpoint of the message's recipient, in our opinion, is the only one that allows us to account for what actually happens in an interaction.** Whether we deem a statement to be an insult or not, we can hardly argue that there is verbal violence if its recipient does not feel insulted. Tolerance for confrontation varies greatly among individuals and communities, hence our only analytical choice is to consider that **it is the reaction to such acts by an addressee that "constructs" the threatening act**, even though the act's intention often aligns with its perception".

However, adopting this analytical stance is not straightforward, as we only have access to speech acts through reported speech. To categorise speech acts as condemnation acts, we would need access to the original interaction to observe if there were any reactions such as

denial or avoidance—which would suggest that these acts were perceived as face threats (Laforest & Vincent, 2004). In addition to our limited access to the interaction, analysing reported speech acts can raise questions about the reliability of reported speech because speakers necessarily, consciously or not, alter the speech they report. The representation of the parent's speech should not be considered a transparent reproduction of what was said; it reflects the speaker's positioning regarding both the reported speech and its enunciator (Rosier, 1999).

This representation is nonetheless valuable and contributes to the (socio)discursive production achieved in this particular context. The account name, *Parents toxiques*, establishes a thematic unity, i.e., abusive or "toxic" parents; a principle of relevance applies, as well as a communication contract (Charaudeau, 2011). I thus consider that selecting and reporting the parent's speech acts in testimonies submitted for publication is a sufficient indication that the authors have viewed or view these acts as threatening and/or disqualifying. Since the condemnation acts have existed pragmatically, through the effects produced on their recipients (Laforest & Moïse, 2013; Moïse et al., 2019), the analysis is not hindered by the authors' enunciative interventions.

In conclusion, the analysis of reported condemnation acts proposed in this article does not claim to provide access to illocutionary acts, i.e., the speech acts performed by the parent. Rather, this corpus allows us to investigate perlocutionary acts, that is, the effects produced on the recipient—in this case, the author. My aim is thus to study the speech acts received as violent, considering that their representation in the *Parents toxiques* testimonies indicates a high degree of performativity and memorability.

3. Between Insult, Contempt, and Reproach: Discursive Modalities of the Condemnation of Being

Based on the analytical framework provided in Sections 1 and 2, the condemnation acts attributed to abusive parents were extracted and divided into two categories: condemnation acts of doing and condemnation acts of being—the latter being our focus. The condemnation acts of being were then categorised, both in terms of pragmatic act (e.g., insult) and topic (e.g., lack of intelligence). At this stage, it became evident that these acts do not function in isolation but interact and overlap, producing layered pragmatic effects; this prevents the discrete coding of each act, as they can simultaneously serve multiple functions. To characterise the power dynamics underlying the condemnation of being, this section will therefore employ a multilayered approach, combining enunciative, argumentative, and pragmatic microanalyses. We will scrutinise three condemnation acts of being present in the corpus: insult, contempt, and reproach. The objective is to examine how the condemnation of being operates in situated contexts and to highlight the discursive devices that underlie it.

3.1. Contempt for the Other: Essentialisation and Downgrading

Composed of *més*, 'bad' and *priser*, 'to estimate', the French verb *mépriser* ('be contemptuous of, to disdain') means "to assign no worth or a derisory worth to a being or a thing" (Bernard Barbeau & Moïse, 2020, §1). Among the four destructive behaviours identified by J. M. Gottman et al. (2019, p. 1213), it is considered the most damaging to relationships: "The contemptuous partner speaks from a place of superiority, and the recipient of contempt feels belittled, put down, and abused. Contempt is expressed in many ways, such as name calling, mockery, sarcasm, and negative comparisons ('You're just like your mother')".

According to Bernard Barbeau and Moïse (2020, 2023), contempt becomes destructive when it is used to gain power over others and to establish a relationship of domination by demeaning them in order to enhance one's own worth. Koselak (2005, §35) also emphasises the relationship of verticality and superiority between the contemptuous subject and

the object of their contempt: *mépriser* is *"to place a (human) object beneath oneself*, which implies being *above them*"—that is, to feel superior towards "those who are too weak (low, small) to deserve any consideration". Contempt thus relies on normative thinking, particularly norms to which the subject adheres: since the contemptuous subject projects their judgement onto their addressee, placing them in a "category of bad objects", they are engaged as a "cognitive subject, source or controller of the judgements expressed" (Koselak, 2005, §35). The negative estimation of the addressee's worth by the contemptuous subject justifies their contempt towards them, granting them the "'right' to disrespect them" (Baider, 2020, §3).

In the condemnation acts reported by the speakers, we observe numerous pejorative qualifications (Laforest & Vincent, 2004) that focus on lack of intelligence and/or competence:

> [4] He often used to say to us, **"You really are good for nothing! You'll never get anywhere in life. Maybe you can clean the toilet at my job!"** [. . .] However, to this day, when I do something (I graduated with a Bac+2 [=two years of higher education] in 2014), and I've been cosplaying for several years, **it is worthless in his eyes. I'm nothing. I'm still a good for nothing to him because I'm unemployed**. . .
>
> *Il nous disait* ***souvent "t'es vraiment bon.ne à rien! T'arriveras jamais à quoi que ce soit dans la vie.. Tu pourras peut-être nettoyer les chiottes à mon boulot!"*** *[. . .]* Cependant, encore aujourd'hui, quand je réalise quelque chose (j'ai validé un bac+2 en 2014), et je fais du cosplay depuis plusieurs années, **ça n'a aucune valeur à ses yeux. Je ne suis rien. Je reste une bonne à rien pour lui parce que je suis sans emploi**. . . (Mollie)
>
> [5] He'd make me work on my maths for hours and when I didn't understand, **he'd call me a "moron"** and shake me by the arm—always that arm. . .. [. . .] More generally, **it was me who wasn't good enough**.
>
> *Il me faisait travailler mes maths pendant des heures et lorsque je ne comprenais pas,* ***il m'insultait de "conne"*** *et me secouait par le bras—toujours ce bras. . . [. . .] Plus généralement,* ***c'était moi qui n'était pas assez bien****. (Chiara)*
>
> [6] He took advantage of a moment when I was burnt out at work **to tell me that I was a good for nothing anyway etc**. . .
>
> *Il a profité d'un moment où j'ai fait un burn out dans mon boulot pour* ***me dire que de toute façon je ne suis qu'un bon à rien etc****. . . (Matthias)*
>
> [7] "We wanted a boy, you were our last attempt". **Hearing night and day that we were incapable, "I should've cut my balls off when I see these sub-sh.".** destroyed my school education.
>
> *"On voulait un garçon, t'étais notre dernière tentative".* ***Entendre nuit et jour que nous étions des incapables, "J'aurai dû me couper les couilles quand je vois ces sous-m.. "*** *a détruit ma scolarité. (Jane)*
>
> [8] Then I lived with my mother, who taught us with **"You're retarded", "What have I done to produce such morons?!"**, "Your father doesn't love you, he doesn't give a fuck about his children".
>
> *J'ai vécu par la suite avec ma mère qui nous éduquait à coup de* ***"Tu es retardée", "Qu'est-ce que j'ai fait pour pondre des abrutis pareils ?!"****, "Votre père ne vous aime pas, il en a rien à foutre de ses enfants". (Gabrielle)*

Negative axiological terms (i.e., conveying the speaker's evaluative judgement, Kerbrat-Orecchioni, 1997) are used to describe the addressee: *bon/bonne à rien*, *retardée*, *incapables*, *sous-m[erdes]*, *abrutis pareils*, *conne*. These are ontotypical insults—a type of

insult that targets a person "in their very being" (Rosier, 2009, p. 68). They are based on "supposedly ontological characteristics of the individual" (Ernotte & Rosier, 2004, p. 35), for instance, being slow, ugly, or clumsy. Examples include being stupid in [5] (*conne*) and [8] (*retardée*), with a pathologisation of stupidity added in the latter case. These traits are exacerbated by the vulgarity and excessiveness of the terms used—which makes them lean towards "essentialist" rather than "situational" insults (Ernotte & Rosier, 2004). Indeed, a high degree of condemnation can be observed: the lack of intelligence and/or competence is not confined to a specific situation or domain but suggests a generalised lack of worth.

In [7], the noun *incapables* is not followed by any postmodifications that would restrict its scope by applying it to a particular "doing" and/or a specific space and time (e.g., incapable of doing X). The expression *[B]on/bonne à rien* ('good for nothing') in [4] and [6] extends the lack of ability to everything, with no exception. In [5], a general lack of worth is reproached without specifying the domain (*n'était pas assez bien*, 'was not good enough'), which tends towards an essentialisation of the recipient, as a person of lesser worth. In [7], where *sous-m*. . . ('sub-sh. . .') evidently suggests the vulgar term *merde* ('shit'), the condemnation extends to the speaker's intrinsic worth, as a person. The assault on their being is profound and emphatic, downgrading them to a status lower than the category *merde*.

The copula *être* ('to be') takes on a value of general truth in these acts. Indeed, it is not limited to a particular space and time, and it attributes a predicate that expresses a high degree of incompetence and/or unintelligence to the subject referent. In [4], the condemnation based on lack of intelligence and competence explicitly extends into the future with a prediction (*T'arriveras jamais à quoi que ce soit dans la vie. . . Tu pourras peut-être nettoyer les chiottes à mon boulot!*). The condemnation is at its highest degree here, expressing an absolute and irremediable lack of competence that applies to everything (*quoi que ce soit*, 'anything at all') and to the entirety of the addressee's existence (*jamais*, 'never'; *dans la vie,'*in life').

In this same excerpt, contempt is particularly expressed in the statement that follows (*Tu pourras peut-être nettoyer les chiottes à mon boulot!*). Contempt is conveyed through the indirect denigration of third parties, the cleaning staff. The enunciator[10] (i.e., the speaker's father) does not refer to cleaning just any premises, but specifically mentions toilets—a "low" place associated with human waste. The choice of a vulgar term (*chiottes*) rather than the more neutral term *toilettes* reinforces the negative axiology and thus the placement of this professional activity beneath the enunciator. A hierarchy is established by the father here, assigning low status to both his daughter and the cleaning staff while positioning himself in a higher position as someone who has achieved an acceptable or valorised status. This positioning is closely linked to the value system to which the enunciator adheres—the idea that cleaning is an inferior profession is taken for granted.

Moreover, the reported ontotypical insults do not seem to stem from an isolated act due to an accidental loss of control on the part of the parent. Indeed, the authors indicate a repeated or even persistent occurrence of certain speech acts (e.g., *souvent*, 'often'; *nuit et jour*, 'day and night') and the accumulation of contemptuous acts in the parent's discourse. Repetition is an aggravating factor in verbal violence because, once repeated, an act can no longer be seen as a mere outburst (Vincent, 2013). The pragmatic effects of condemnation are amplified by the parent's "conscious adherence" to "both the form and content" (Vincent, 2013, p. 41) of the repeated acts.

The ontotypical insults regarding lack of intelligence and competence discussed in this section constitute the majority of the insults identified in the corpus, and they all fall under the contempt category. According to Ernotte and Rosier (2004), ontotypes are not always perceived as insults by the interactants: they are less salient than sexotypes and

ethnotypes (i.e., insults based on the target's gender and ethnicity, respectively) and are more socially acceptable. However, their strong representation in the speech acts reported by the speakers suggests that they can, indeed, cause violent effects. In the following section, we delve deeper into the indirect expression of contempt and its implications in terms of performativity.

3.2. Performativity of the Indirect Condemnation Act

Analysts of verbal violence have described contempt as an indirect speech act (Moïse et al., 2019; Bernard Barbeau & Moïse, 2020, 2023) as it relies on "implicit and implied forms aimed at inducing negative emotions in others, such as feelings of shame and guilt" (Bernard Barbeau & Moïse, 2023, §4). Because it is expressed through inferences (Baider, 2020) and does not have canonical forms, objectifying it is not straightforward in interaction[11]. As noted previously, this speech act aims to assert superiority and power over the other (Koselak, 2005; Bernard Barbeau & Moïse, 2020, 2023), "favoring the silence of the addressee (ostracism)" (Baider, 2020, p. 3). The asymmetry of the relationship constitutes an aggravating factor according to Bernard Barbeau and Moïse (2023, §4), as it may hinder the possibility of responding or taking counteraction:

> "In an asymmetrical relationship, verbal violence realised through indirect speech acts is more pronounced when the person in the higher position (the parent, teacher, superior, etc) holds symbolic power. It is challenging for the person in the lower position (the child, student, employee, etc) to respond without fearing potential consequences, whether affective or professional".

Moreover, contempt is particularly violent when it comes from esteemed individuals (Bernard Barbeau & Moïse, 2023, §3). The interaction between parent and child thus presents two aggravating factors: asymmetrical positions within an authority relationship, with the parent in a higher position and the child in a lower position, and affective proximity between the interactants.

To illustrate the indirect nature of contempt, let us examine an act reported by Julie. Unlike the examples in the previous section, which focused on the lack of intelligence and competence, this extract involves the denigration of physical appearance:

> [9] Then comments about my appearance: "Cover your legs, they're too skinny". "You were so ugly as a baby". And in the middle of my teenage crisis and so of my insecurities, I hear from her mouth, **"You may not be beautiful, but you're intelligent"** (I was a good student).
>
> *Puis des remarques sur mon physique: "cache tes jambes, elles sont trop maigres". "Tu étais si laide bébé". Et en pleine crise d'ado et donc de complexes, j'entends de sa bouche* ***"Tu n'es peut-être pas belle mais tu es intelligente"*** *(j'étais bonne élève). (Julie)*

In the highlighted segment, we identify the two stages of concessions described by Doury and Kerbrat-Orecchioni (2011, §22 citing Moeschler & De Spengler, 1981, 1982; Morel, 1996):

- "First, an argument p is put forward for a conclusion r. The speaker may either express agreement with p (acknowledging its relevance or truth value; Moeschler & De Spengler, 1981, p. 101) or, more modestly, suspend their judgement (Léard & Lagacé, 1985, pp. 14–15); in either case, p is not contested.
- p is followed by another argument q for a non-r conclusion; q is typically introduced by an oppositional connector (typically "mais" [in French, i.e., 'but']). This second argument is presented as outweighing the first (the concessive movement thus leads to the conclusion of non-r) [. . .]".

In the first independent clause, the speaker concedes the truth value of the predicative relation <*you—not be beautiful*> before introducing a second assertion based on the same topic (*tu*, 'you') with the oppositional connector *mais* ('but'). Although they address different aspects—the addressee's beauty and intelligence—these two assertions can be seen as opposing arguments in response to an enquiry about the addressee's worth. At first glance, this concession appears to lead to a compliment, as suggested by the use of the positive axiological term *intelligente*, thereby valorising the recipient: the concessive movement tends towards an invalidation of the conclusion implied by the first assertion (*tu n'es peut-être pas belle*, 'you may not be beautiful'), in favour of an alternative conclusion.

Concessions are fundamentally dialogical, as conceded utterances echo previous utterances (Doury & Kerbrat-Orecchioni, 2011). In our case, the concession's dialogical nature is marked by *peut-être* ('maybe'), which signals that the enunciator (i.e., the speaker's mother) is reacting to a prior utterance. Interestingly, the speaker's mother takes minimal enunciative responsibility for this denigrating utterance. This is indicated by the use of the concessive marker, the absence of *je* ('I'), and the foregrounding of a valorising assertion (*mais tu es intelligente*, 'but you're intelligent'). Yet, the other utterances attributed to her by the speaker—*cache tes jambes, elles sont trop maigres; Tu étais si laide bébé*—suggest that she is the enunciative source of these utterances.

This concession presents an unusual placement of contentious information, namely, the negation of the addressee's beauty. If we switch *intelligente* and *pas belle*, we obtain a prototypical concession: *Tu es peut-être intelligente, mais tu n'es pas belle* ('You may be intelligent, but you're not beautiful'). In the reformulated utterance, the speaker would first demonstrate—or feign—cooperation with her addressee and then introduce an argument that diverges from agreement between the interactants and outweighs the first. The element presented as non-contentious (i.e., her intelligence) would flatter the addressee. In contrast, in [9], the conceded utterance is face-threatening and disqualifying.

As a result, this concession pertains to *polirudeness*[12] (Kerbrat-Orecchioni, 2010, §32), which covers "utterances that appear to be face-flattering acts (therefore 'polite' utterances) but underneath which lies a face-threatening act (the prototypical case being what can be called 'backhanded compliments' such as 'Your hair looks nice today')". Because of its ambivalence, polirudeness contributes to indirect verbal violence (Moïse & Oprea, 2015). Here, contempt is constructed by presenting the recipient's lack of worth as unsurprising: by placing denigration in the conceded part, the enunciator presents its content as presupposed and shared information.

In essence, this act of contempt achieves an ontological assertion of power over the addressee, as the enunciator assumes and exercises the right to define the addressee's very being. The combination of two acts with conflicting illocutionary values makes it challenging to objectivise the disqualifying act, thereby hindering the possibility of a response. The condemnation is constructed as a shared and self-evident fact rather than a judgement issued by the enunciator: the discrepancy between the recipient's physical appearance and the beauty standards of the contemptuous subject is not treated as a mere difference but as a failure to meet a universal norm—which legitimises placing them "beneath oneself" (Koselak, 2005).

3.3. Stratification of the Condemnation of Being

In the previous section, we observed how contempt, through its indirect actualisation, can be difficult to objectify in interactions. In this section, we further explore its placement in the enunciative background, which is particularly evident in excerpt [10] below. The children are denigrated with the pejorative qualification *des abrutis pareils* ('such morons'); the negative axiology is reinforced by the verb *pondre* ('to lay [an egg]', colloquially 'to

produce'), which takes on a vulgar connotation by dehumanising birth. Both the pejorative qualification and the verb contribute to acts of insult and contempt.

> [10] Then I lived with my mother, who taught us with "You're retarded", **"What have I done to produce such morons?!"**, "Your father doesn't love you, he doesn't give a fuck about his children".
>
> *J'ai vécu par la suite avec ma mère qui nous éduquait à coup de "Tu es retardée", **"Qu'est-ce que j'ai fait pour pondre des abrutis pareils ?!"**, "Votre père ne vous aime pas, il en a rien à foutre de ses enfants". (Gabrielle)*

Interestingly, children's categorisation is presupposed. Indeed, the denigration based on lack of intelligence and/or competence and the ensuing contempt are situated in the enunciative background, within the prepositional phrase *pour pondre des abrutis pareils*. We observe a nominalisation: the enunciator presents the process—and therefore the categorisation of the addressee—as already established through the use of an infinitive clause embedded within a larger syntactic structure. This syntactic process illustrates what occurs at the pragmatic level: acts of insult and contempt are framed as shared knowledge and serve as the foundation for another indirect speech act, guilt induction. Within such stratification, it is the harm experienced by the parent—rather than the ontotypical insult or contempt—that is brought to the enunciative foreground.

In this regard, we can observe a delocution of the addressee, i.e., speaking about someone who is present rather than addressing them directly: there is no second-person pronoun, and the expressive phrase *qu'est-ce que j'ai pu faire pour*. . . ('what could have I done to. . .') neither has a direct recipient nor invites a response. The adjective *pareils* ('such') intensifies the condemnation, suggesting that the noun *abrutis* ('morons') alone is insufficient to convey the children's lack of intelligence and competence. Placing condemnation in the presupposed, as non-contentious information and therefore unsurprising to the addressee, tends towards the normalisation of contempt.

This stratification can also be observed in the prototypical feature of abusive parent-child interactions identified by Van Hooland (2005, 2008): renaming the child with derogatory terms of address, which also frames the condemnation as shared knowledge. As noted earlier by Bernard Barbeau and Moïse (2020, 2023), the performativity of condemnation is amplified by the indirectness of speech acts and the underlying authority relationship: by acknowledging that a pejorative qualification refers to them, and even more so by implicitly accepting it (e.g., by responding to it), the interactant may appear to co-construct their subordinate position and, to some extent, the condemnation itself. However, attempting to refute the adequacy between the pejorative qualification and oneself would present risks of retaliation, such as physical violence or affective consequences (Bernard Barbeau & Moïse, 2020, 2023).

I argue that the stratification of condemnation extends to all condemnation acts of being and plays a central role in power dynamics. Indeed, the statement *T'es vraiment bonne à rien!* ('You really are good for nothing!', Mollie) not only achieves acts of insult and contempt, but also serves as a reproach. Figure 3 below illustrates this stratification, with a bidirectional dynamic between the (perceived) ontological properties of the target and the speaker's beliefs and needs. As seen in the microanalysis, framing the condemnation of being as evident, shared information functions as an aggravating device. For instance, in *Qu'est-ce que j'ai fait pour pondre des abrutis pareils?!*, guilt induction is more salient as the prejudice experienced by the parent is foregrounded; the insult is presupposed (as opposed to the more direct *Vous êtes des abrutis*), which normalises contempt and reinforces the subordinate position of the addressee.

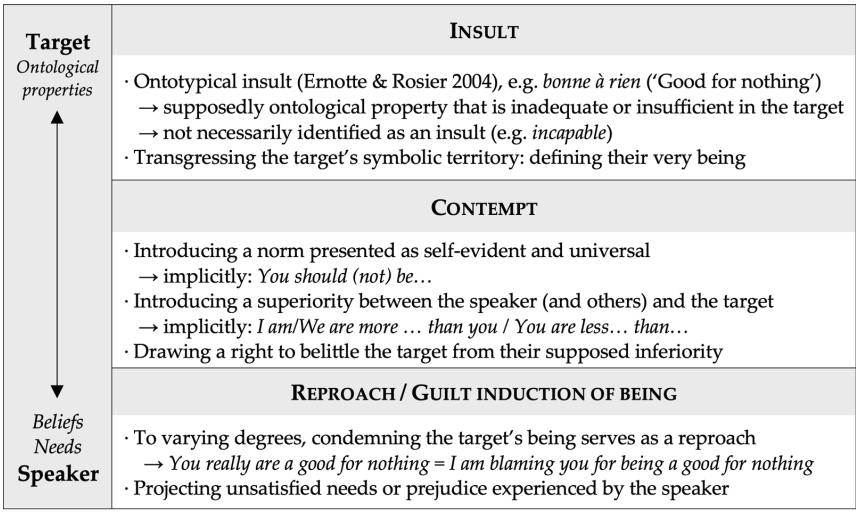


Figure 3. Stratification of the Condemnation of Being.

4. Constructing the Being at Fault: Negation of Otherness and Disruption of the Authority Relationship

At this stage of our reflection, we can agree that condemning someone's being equates to subordinating them to oneself: the contemptuous subject places the other beneath themselves by assigning them a lower position while simultaneously positioning themselves above them. This subordination involves apprehending the addressee "through oneself", as the subject perceives them through the lens of their own value system. Otherness is threatened, both by rejection—as being inadequate—and by the failure to recognise and treat it as distinct from oneself. In the final section, I therefore explore the reasons invoked for constructing the being as at fault and the resulting disruption of the authority relationship.

4.1. Reproach and Guilt Induction: From Otherness to Oneself

Reproach is a direct speech act by which a speaker expresses their "disapproval of the being and/or doing" (Moïse et al., 2019, p. 85) of a target who is their addressee[13]. It often displays a shift from condemning one's doing (i.e., actions, behaviours) to condemning one's being (Moïse et al., 2019), where the being is condemned in order to reproach a specific behaviour (Laforest & Vincent, 2004). Like contempt, it stems from a comparison with a norm, highlighting the insufficiency or inadequacy of certain behaviours or characteristics in contrast to what is expected. Reproach thus signals what is expected of the interactant—whether these expectations are consciously acknowledged or not—and an ontological assertion of power over them: the behaviour deemed at fault "serves as proof of what is not and should be, of what the person is not and should be" (Moïse et al., 2019, p. 85). Guilt induction of the inadequate behaviour's author often underpins the act of reproach. It is an indirect speech act that "aims to place the other in debt by invoking a lack of recognition; it demands reparation, which prompts the guilt-inducted party to conform to the wishes of their guilt inducer" (Neuburger, 2008, cited by Moïse et al., 2019, p. 89).

The distinction between direct and indirect speech acts, although useful in accounting for the insidious nature of guilt induction, does not allow us to easily differentiate between these two acts. Having no direct access to the child's reactions or the parent's intentions, I have no analytical choice but to consider it impossible to determine whether a given reproach is coupled with guilt induction or not. However, as observed in Section 2.2, the

situation of utterance suggests that the speakers have indeed considered or are considering the acts they report as inappropriate and/or harmful. I shall then follow the approach adopted by Laforest and Vincent (2004, pp. 64–65), who identify five types of shortcomings reproached to the target through the use of pejorative qualifications (i.e., lack of strength or courage, lack of experience or maturity, lack of intelligence, lack of consideration or respect for others, and lack of respectability). Building on their approach, I identify the reasons invoked by the parent when the condemnation targets the being and the underlying relationship to otherness.

4.2. Reasons for the Guilt Induction of Being

Three interrelated reasons for the guilt induction of being are invoked: lacking worth, not being satisfying and/or lovable as a child, and existing as one is and/or being born (see Figure 4). These reasons are often combined, especially in the most violent acts.

1	Lacking worth
	[11] He often used to say to us, **"You really are good for nothing!** *Il nous disait souvent "**t'es vraiment bon.ne à rien !** [...]" (Mollie)*
	[12] She was drooling, **shouting that I was just a slut, a real bitch, vile and ungrateful**. *Elle bavait, **criant que j'étais qu'une salope, une vraie connasse, immonde et ingrate**. (Gabrielle)*
2	Not being satisfying and/or lovable as a child
	[13] I've lost count of the number of times **I've heard that I'm my father's daughter...** without really understanding why, unfortunately. *Je ne compte plus les fois où **j'ai entendu que j'étais la digne fille de mon père**… sans vraiment comprendre pourquoi malheureusement. (Silvia)*
	[14] [...] he used to pick on me, forcing me to leave his house and **telling me he hated me and never wanted to see me again**. [...] *il s'acharnait sur moi, m'obligeant alors à partir de chez lui **en me disant qu'il me détestait et qu'il ne voulait plus jamais me voir**. (Albane)*
3	Existing as one is and/or being born
	[15] Then I lived with my mother, who taught us with "You're retarded", **"What have I done to produce such morons?!"** […] *J'ai vécu par la suite avec ma mère qui nous éduquait à coup de "Tu es retardée", **"Qu'est-ce que j'ai fait pour pondre des abrutis pareils ?!"** […]. (Gabrielle)*
	[16] **"We wanted a boy, you were our last attempt".** Hearing night and day that we were incapable, **"I should've cut my balls off when I see these sub-sh.."** destroyed my school education. ***"On voulait un garçon, t'étais notre dernière tentative"**. Entendre nuit et jour que nous étions des incapables, "**J'aurai dû me couper les couilles quand je vois ces sous-m.."** a détruit ma scolarité. (Jane)*
	[17] I'm an unwanted child, born at the end of the '80s when my mother was barely 21. [...]. As far back as I can remember, **she always told me that I was the biggest mistake of her life**. *Je suis un enfant non désiré, né à la fin des années 80 alors que ma mère avait à peine 21 ans.* […]. *Aussi loin que je m'en souvienne, **elle m'a toujours dit que j'étais la plus grosse erreur de sa vie**. (Matthias)*

Figure 4. Reasons for the Guilt Induction of Being.

The first category includes condemnation acts such as those analysed in Section 3.1 (e.g., [11]); they contain contemptuous ontotypical insults that portray the child as inadequate and insufficient by reproaching a lack of worth. This lack of worth reproached to the target can consist of a lack of intelligence or competence, but also a lack of recognition towards the parent, as seen in [12].

In the second category, guilt induction relates to not being satisfactory or lovable as a child; a lack of worth is implied. Indeed, the object of guilt induction is not explicitly stated in [13] (*la digne fille de mon père*) or in [14]. However, the latter extract features an explicit

and emphatic negation of love towards the addressee, conveyed by the verb *détester* and the high-degree adverbial phrase *plus jamais* ('never ever'). It is noteworthy (as observed earlier regarding contempt) that the shortcomings reproached here are not isolated and modifiable behaviours—the object of guilt induction is the target in their being and in their entirety.

At its extreme, guilt induction of being leads to a third category: the target is blamed for existing as they are and/or for being born as they are. It targets the fact of existing as an insufficiently intelligent and competent person in [15], and as a girl rather than a boy in [16]; in [17], guilt induction is based on being born despite not being wanted. In such cases, all three reasons for guilt induction are combined: by presenting the child's birth and existence as detrimental to the parent, these reproaches of existence and/or birth also imply that the recipient is not satisfactory and lovable as a child, and lacks worth. The transgression of the other's symbolic territory is absolute, as their right to exist is attacked, and this seems to occur repeatedly (*entendre* ***nuit et jour*** ['day and night'] *que*. . .; *elle m'a* ***toujours*** ['always'] *dit que*. . .; *ma mère qui nous éduquait* ***à coup de*** ['constantly, repeatedly']. . .). By denying the right to exist as one is and expressing the harm caused by their very being, these guilt-inducing acts tend towards a symbolic destruction of the target. They fulfil the three criteria of direct hate speech identified by Lorenzi Bailly and Moïse (2021, p. 12): the use of condemnation acts, reliance on a pathemic dimension, and the presence of markers that negate otherness.

The harm caused by the target is salient in the guilt induction of being: the three reasons invoked pertain to the child being detrimental to the parent, negatively affecting their life, and/or constituting a burden. The condemnation of being places the addressee in debt, constructing and assigning them a diffuse fault. Guilt induction is accompanied—to varying degrees—by victimisation, particularly manifested through the expression of harm and regret (e.g., *Qu'est-ce que j'ai fait pour*. . ., 'What have I done to. . .'; *J'aurai[s] dû*. . ., 'I should've. . .') and emphatic modifiers (e.g., *des abrutis* ***pareils,*** 'such morons'; *ces* ***sous****-m*[*erdes*], 'sub-sh[its]'; *la* ***plus grosse*** *erreur* ***de ma vie,*** 'the biggest mistake in my life'). The parent thus legitimises the condemnation of being, presenting it as a reaction to an initial fault. This is evident in example [12], cited in the previous section and reproduced below with more context. The speaker recounts a scene in which verbal violence occurs after physical violence:

> [12'] I'd discovered that my mother was going through all my messages and that she'd done the same with my sister's mailbox and she was really angry about what she'd found there: messages from my sister saying we were unhappy./**So her reaction to that was to punch me in the face. Which stunned me for a moment.** My brother held her back or she'd have jumped on me. She was drooling, **shouting that I was just a slut, a real bitch, vile and ungrateful**.
>
> *J'avais découvert que ma mère fouillait tous mes messages et qu'elle avait fait de même avec la messagerie de ma sœur et elle était vraiment en colère de ce qu'elle y avait trouvé: des messages de ma sœur disant que nous étions malheureux./****Alors la réaction qu'elle a eu face à ça et de me coller une droite en pleine face. Me sonnant pendant un instant****. Mon frère l'a retenu sinon elle me sautait dessus. Elle bavait,* ***criant que j'étais qu'une salope, une vraie connasse, immonde et ingrate****. (Gabrielle)*

This is noteworthy because physical violence, being "the last resort to make oneself heard", typically erupts after verbal violence (Moïse et al., 2019, p. 13)[14]. The acute verbal violence (Moïse et al., 2019) at work here, involving a direct condemnation act of being (i.e., an insult), appears to legitimise the preceding physical violence. Indeed, the lack of worth reproached by the parent, expressed through pejorative qualifications (*une salope, une vraie connasse, immonde et ingrate*, 'a slut, a real bitch, vile and ungrateful') consists in a lack of

recognition towards the parent as well as a lack of respectability. These shortcomings are presented as an affront and legitimise physical violence.

In conclusion, guilt induction of being intertwines the target's intrinsic worth with their affective bond to the parent. The reasons invoked for the condemnation of being frame the child as an object through and within the parent's perspective, as they do not align with the latter's expectations—both in terms of beliefs and needs.

4.3. Towards a Disruption of Interactional Dynamics

We must now address the question asked in the introduction: is the authority relationship that characterises the parent-child interaction disrupted in and by condemnation acts of being, and if so, how?

An essential characteristic of the authority relationship is that the interactants mutually recognise each other in their high and low positions, and that this relationship is co-constructed (Moïse et al., 2019). However, resorting to speech acts that undermine the child's face, with disqualifications that attack self-esteem across different spheres (i.e., social sphere, intimate sphere, and values—Moïse et al., 2019, p. 81) departs from a recognition of the one in the lower position, and thus a co-construction of the relationship. Throughout the analysis of condemnation acts of being, it has become apparent that the position of guarantor of the interaction, expected in an authority relationship, acquires a sense of superiority over the other. It materialises in an ontological assertion of power over the other, by reaching their worth and essence. As discussed in Section 3, the categorisation and attribution of disqualifying traits are sometimes situated within the presupposed, which frames them as shared knowledge. In such cases, the child's low position is not a protective and supportive one but, rather, becomes a site of attacks on their being with latent and reactivable disqualifications. The asymmetry of positions is thus exacerbated in and by condemnation acts of being, leading to a relationship of domination. The parent becomes an all-powerful subject, treating the other as an object through the lens of their values, expectations, and needs; the child is objectified through the implicit or explicit comparisons to what they should be that underly contempt and guilt induction.

While the analysed condemnation acts of being display an assertion of power over the other, they also appear to involve a symmetrisation of positions. Through guilt-inducing acts, the parent places the child in a position of responsibility and/or symbolic debt that may be inappropriate, especially when it concerns reasons beyond the child's control. This disruption is regularly highlighted in the fragments of interaction represented in the testimonies; in the extracts below, for instance, the young Jane and Matthias criticise their assignment to a position where they must, like peers, advise and take charge of their parent.

> [18] But being your mother's shrink in the evening, with her nose in her bottle, gives a bitter taste of life as early as 6 years old, just as she experienced it
>
> *Mais être la psy de sa mère le soir, le nez dans sa bouteille, donne dès ses 6 ans un goût amer de la vie, autant qu'elle le vivait. (Jane)*
>
> [19] When I was a kid, I quickly became the man of the house, at least when she was single. Which means I became her confidant, sharing all her troubles. One evening when I was about 6 or 7, we were sitting on the sofa, she took a handful of pills and told me "goodbye". Obviously I panicked, and when I tried to call 911 she told me off, because "it was just a joke to see if I loved her".
>
> *Quand j'étais petit, je suis rapidement devenu l'homme de la maison, en tout cas dans les moments où elle était célibataire. Ce qui signifie que je suis devenu son confident en partageant tous ses malheurs. Un soir, vers 6 7 ans, nous étions assis sur le canapé et elle a prit une poignée de cachets en me disant "Adieu". Forcément j'ai paniqué et lorsque*

j'ai voulu appeler les pompiers elle m'a disputé, car « ce n'était qu'une blague pour voir si je l'aimais ». (Matthias)

In sum, the authority relationship is disrupted by a dual movement within the condemnation acts of being: the exacerbation of asymmetry, and the symmetrisation of interactional positions.

5. Conclusions

This article aimed to analyse the treatment of otherness underlying the condemnation acts of being reported in anonymous testimonies published on the Instagram account *Parents toxiques*. After outlining the study's methodological and analytical framework, I examined how one's being is constructed as being at fault. The linguistic and discursive devices identified through microanalyses, along with their implications in terms of interactional dynamics, are summarised in Figure 5 below. It breaks down how the condemnation of being is enacted and the treatment of otherness involved by such condemnation.

1. Predicative level: ontological assertion of power over the other
· Present tense with general truth value · State predicates, notably with the copula *be* · No adverbs/adverbial phrases that restrict the process to a particular time and space · Intensifiers: high-degree and restrictive adverbs, e.g. *never*, *nothing*, *always*, *only*, *just* · Future tense with predictive value (extension of the condemnation) · No enunciative responsibility for the condemnation with *I* · No modalisation of the condemnation (no expressions such as *In my view/opinion…*)
2. Argumentative and pragmatic levels: subordination of the other to oneself
· Insertion of negative axiology: pejorative qualifications, emphatic modifiers, vulgar language · Framing the condemnation as shared knowledge (presupposition) · Reliance on norms presented as self-evident and universal · Legitimisation of the condemnation: framed as a reaction to an initial fault or act of violence · Pathemic effects that tend towards victimisation · Face undermining: denigration; combination of pragmatic acts with conflicting values (polirudeness / mock politeness); delocution
3. Interactional and relational levels: disruption of the authority relationship
· Conflation of the other's intrinsic worth with the speaker's beliefs and needs · Exacerbation of the asymmetry → relationship of domination · Symmetrisation of the interactional positions → the participant in the higher position does not assume the role of guarantor of the interaction; responsibility or even symbolic debt is assigned to the participant in the lower position

Figure 5. Modalities of the Condemnation of Being.

Based on microanalyses of the condemnation acts of being reported in our sample of *Parents toxiques* testimonies, we can draw the following conclusions. (i) The ontological assertion of power over the other is constructed from the predicative level, with processes that concern the recipient's being in its entirety and present the condemnation as an objective reality. (ii) The condemnation of being draws its pragmatic force from its legitimisation—by relying on norms presented as self-evident and universal, and by highlighting the harm caused by the other. (iii) As a speaker, constructing the other's being as at fault involves, to varying degrees, essentialising and downgrading them, as well as conflating their intrinsic worth with one's beliefs and needs; these two objectifying processes hinder the recognition

of otherness as distinct from oneself, which is necessary to acknowledge the other as a subject (Moïse, 2020).

These findings should be considered in perspective. Given the ethical and methodological challenges of accessing and analysing authentic abusive interactions highlighted in Section 1—namely, the difficulty of categorising them as such—we used the speech acts reported in the *Parents toxiques* testimonies as an entry point to understand how verbal abuse may be committed and received. As discussed in Section 2, this perspective is inherently limited: we access condemnation acts as received and remembered by the authors, not as they occurred in real-time interactions. I argue that the authors' situated viewpoints and the means employed to represent their experience should not be seen as biases but as a research focus. In other words, the enunciative and argumentative interventions of the authors are an inherent part of what these anonymous online testimonies allow us to observe. This approach to experiential data aims to avoid value judgements on the part of the analyst: we do not question or confirm the testimonies' legitimacy and truthfulness, but treat them as *sociodiscursive traces* that need to be deciphered.

On the one hand, these traces point to how certain parental behaviours were experienced, remembered, and later reflected upon, which gives us privileged access to the long-term effects attributed to behaviours perceived as abusive or "toxic". Notably, our reception-based findings resonate with experimental psychology studies conducted in controlled environments that identified criticism and contempt as damaging to relationships (J. M. Gottman et al., 2019). If we view the condemnation of being as a means of domination, both in terms of attitude to otherness and interactional dynamics, the notion of condemnation of being—along with its descriptors—provides an effective framework that can be applied to reports and direct observations. It can help various professionals identify and assess transgressions and/or dysfunctions in authority relationships—not only in the family sphere but also, for instance, in the workplace.

On the other hand, studying these traces allows us to characterise what the authors *do* through their testimonies, and therefore to question the social meaning acquired by the anonymous publicisation of their experience. The decontextualisation at work in these testimonies—since we do not know who is testifying, when, and where—should not be treated as missing information but as a discursive mechanism that both the *Parents toxiques* account and the authors engage in (Moreau Raguenes, 2024b). To build on the present study, further research should examine the enunciative, argumentative, and pragmatic materiality of reported speech itself—that is, what the authors achieve when reporting their parents' condemnation acts (e.g., denouncing, invalidating, providing proof).

Finally, the sample of ten testimonies consists of texts that are longer than average on *Parents toxiques*, and may therefore use more linguistic resources than shorter testimonies; our findings will need to be tested against the full dataset of 314 testimonies, which was compiled at a later stage of the project.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of CY Cergy Paris Université on 8 December 2022 (protocol code: 202212-001).

Informed Consent Statement: It was not possible to obtain consent from the authors of the testimonies as these are anonymous testimonies published on a public Instagram account. The account creator invites users to send their testimonies via message; they are then published anonymously, without identifying or naming the author. To further preserve the anonymity of the authors, the publication date and URL of the testimonies analysed in this article were not included. Additionally, potentially identifying information, though rare in the corpus, has been anonymised (e.g., city names,

first names)—such anonymisation was not required in the sample analysed for this article. Analytical precautions have also been taken to protect the authors' integrity: the analysis does not question or confirm the truthfulness of the experience they recount. Based on this, the Research Ethics Committee of CY Cergy Paris Université unanimously approved this research.

Data Availability Statement: The data presented in this article (i.e., extracts from anonymous testimonies) are available at https://www.instagram.com/parentstoxiques/ (accessed on 15 November 2024).

Acknowledgments: I would like to thank Julien Longhi, Laurence Rosier, and Claudine Moïse for their invaluable contributions throughout this research project, as well as the anonymous reviewers for their insightful feedback. This article is part of the research conducted by the international research group *Draine, Haine et rupture sociale: discours et performativité*, which studies hate speech, extremist narratives, violent discourse, and the respective genres associated with them (https://groupedraine.github.io/). This article is a revised and enriched translation of Moreau Raguenes (2024a) "L'être comme fautif. Actes de condamnation, altérité et rapport de places dans l'interaction parent-enfant" originally published in French by EDP Sciences in *SHS Web of Conferences*, Volume 191, article 01009. The translation was prepared by Rose Moreau Raguenes, assisted by generative AI tools (ChatGPT, DeepL). The author wishes to thank EDP Sciences for granting permission as well as the organisers of the 2024 World Congress of French Linguistics.

Conflicts of Interest: The author declares no conflict of interest.

Notes

1 All translations will be mine throughout the article unless specified otherwise

2 The *Parents toxiques* corpus was compiled as part of a doctoral research project supervised by Julien Longhi and Laurence Rosier. The sample was collected for a Master's research project supervised by Claudine Moïse (Moreau Raguenes, 2021).

3 https://www.legifrance.gouv.fr/codes/article_lc/LEGIARTI000038749626 (accessed on 15 June 2024).

4 "Chapitre II: Des atteintes à l'intégrité physique ou psychique de la personne (Articles 222-1 à 222-67)" https://www.legifrance.gouv.fr/codes/section_lc/LEGITEXT000006070719/LEGISCTA000006149827/#LEGISCTA000006149827 (accessed on 15 June 2024).

5 I choose to use the notion of "condemnation act" (Laforest & Moïse, 2013 notably, 'actes de condamnation' in French) rather than face-threatening acts to refer to acts that undermine the addressee's face (Goffman, 1956) and identity. This choice emphasises the disqualification of the target—disqualification not being constitutive of all face-threatening acts (e.g., an injunction not accompanied by disqualifying acts).

6 Of course, these parameters vary depending on the child's age.

7 The URL and publication date of the testimonies will not be indicated for ethical reasons.

8 Emphasis (in bold) is always my addition; in the quotes, italics always come from the cited text. Spelling and punctuation will not be modified in the cited corpus excerpts.

9 In [2], Chiara could be reporting a reproach her father directed to her (e.g., You make me angry), in which case the segment enclosed in quotation marks would be reported speech. Alternatively, she might be using quotation marks to express critical distance towards the causal link represented. In [3], it could be an injunction (e.g., Don't make things worse) reported in free indirect speech, but the boundary between what the parent said and what was understood and internalised by the speaker ("Not to make waves. Not to be an extra burden".) is porous.

10 I use the term *enunciator* because the speech acts analysed are in reported speech: the parent is the enunciative source but does not perform the locutionary act. The term *speaker* always refers to the authors of the testimonies.

11 However, argumentative strategies that allow the narrators to express contempt in interaction can be identified through analysis (see Baider, 2020 for instance).

12 In the English-speaking literature, a similar notion has been proposed—*mock politeness* (Taylor, 2015, for instance).

13 I choose to use the term *reproach* rather than *complaint* because, as pointed out by Laforest (2002, p. 1596) "the meaning of the term 'complaint' is broader than that of the French term 'reproche', [. . .] which refers only to dissatisfaction addressed to the person held to be responsible for deviant behavior".

14 See the model of the escalation of acute verbal violence ("violence verbale fulgurante" in French) in Moïse et al. (2019, p. 13).

References

Austin, J. L. (1962). *How to do things with words*. Harvard University Press.

Ayimpam, S. (2015). Enquêter sur la violence. *Civilisations. Revue internationale d'anthropologie et de sciences humaines*, *64*, 3852. [CrossRef]

Baider, F. (2020). Obscurantisme et complotisme: Le mépris dans les débats en ligne consacrés à la vaccination. *Lidil. Revue de linguistique et de didactique des langues*, *61*, 7652. [CrossRef]

Bernard Barbeau, G., & Moïse, C. (2020). Introduction—Le mépris en discours. *Lidil. Revue de Linguistique et de Didactique des Langues*, *61*. Available online: https://journals.openedition.org/lidil/7264 (accessed on 6 March 2025). [CrossRef]

Bernard Barbeau, G., & Moïse, C. (2023). Mépris. In N. L. Bailly, & C. Moïse (Eds.), *Discours de haine et de radicalisation: Les notions clés* (pp. 277–281). ENS Éditions. [CrossRef]

Charaudeau, P. (2011). Du contrat de communication en général. In *Les médias et l'information. L'impossible transparence du discours* (pp. 49–55). De Boeck Supérieur. Available online: https://www.cairn.info/les-medias-et-l-information--9782804166113-page-49.htm (accessed on 6 March 2025).

Clancy, B. (2011). Complementary perspectives on hedging behaviour in family discourse: The analytical synergy of variational pragmatics and corpus linguistics. *International Journal of Corpus Linguistics*, *16*(3), 371–390. [CrossRef]

Doury, M., & Kerbrat-Orecchioni, C. (2011). La place de l'accord dans l'argumentation polémique: Le cas du débat Sarkozy/Royal (2007). *A Contrario*, *16*(2), 63–87. [CrossRef]

Einarsdóttir, J. (2007). Research with children: Methodological and ethical challenges. *European Early Childhood Education Research Journal*, *15*(2), 197–211. [CrossRef]

Ernotte, P., & Rosier, L. (2004). L'ontotype: Une sous-catégorie pertinente pour classer les insultes? *Langue française*, *144*(1), 35–48. [CrossRef]

Fracchiolla, B., Lorenzi Bailly, N., Moïse, C., & Romain, C. (2023). Violence verbale. In N. L. Bailly, & C. Moïse (Eds.), *Discours de haine et de radicalisation: Les notions clés* (pp. 299–307). ENS Éditions. [CrossRef]

Goffman, E. (1956). *The Presentation of self in everyday life*. Doubleday.

Gottman, J., & Gottman, J. (2017). The natural principles of love. *Journal of Family Theory & Review*, *9*(1), 7–26. [CrossRef]

Gottman, J. M., Cole, C., & Cole, D. L. (2019). Four horsemen in couple and family therapy. In J. L. Lebow, A. L. Chambers, & D. C. Breunlin (Eds.), *Encyclopedia of couple and family therapy* (pp. 1212–1216). Springer. [CrossRef]

Kerbrat-Orecchioni, C. (1997). *L'énonciation: De la subjectivité dans le langage* (3rd ed.). Armand Colin.

Kerbrat-Orecchioni, C. (2010). L'impolitesse en interaction. *Aperçus théoriques et étude de cas.* Lexis. Journal in English Lexicology, *HS 2*, 35–60. [CrossRef]

Kopelman, L. (2000). Children as research subjects: A dilemma. *The Journal of Medicine and Philosophy: A Forum for Bioethics and Philosophy of Medicine*, *25*(6), 745–764. [CrossRef]

Koselak, A. (2005). Mépris/dédain, deux mots pour un même sentiment? Lidil. *Revue de Linguistique et de Didactique des Langues*, *32*, 21–34. [CrossRef]

Kousholt, D., & Juhl, P. (2023). Addressing ethical dilemmas in research with young children and families. Situated ethics in collaborative research. *Human Arenas*, *6*(3), 560–579. [CrossRef]

Labov, W. (1972). *Sociolinguistic patterns*. University of Pennsylvania Press.

Laforest, M. (2002). Scenes of family life: Complaining in everyday conversation. *Journal of Pragmatics, Negation and Disagreement*, *34*(10), 1595–1620. [CrossRef]

Laforest, M., & Moïse, C. (2013). Entre reproche et insulte, comment définir les actes de condamnation? In B. Fracchiolla, C. Moïse, C. Romain, & N. Auger (Eds.), *Violences verbales. Analyses, enjeux et perspectives* (pp. 85–105). Presses universitaires de Rennes. Available online: https://hal.archives-ouvertes.fr/hal-01969711 (accessed on 6 March 2025).

Laforest, M., & Vincent, D. (2004). La qualification péjorative dans tous ses états. *Langue Française*, *144*(1), 59–81. [CrossRef]

Laforest, M., & Vincent, D. (2006). *Les interactions asymétriques*. Ed. Nota Bene.

Léard, J.-M., & Lagacé, M. F. (1985). Concession, restriction et opposition: L'apport du québécois à la description des connecteurs français. *Revue québécoise de linguistique*, *1*(15), 11–49. Available online: https://www.erudit.org/fr/revues/rql/1985-v15-n1-rql2925/602548ar/ (accessed on 6 March 2025). [CrossRef]

Longhi, J. (2013). Essai de caractérisation du tweet politique. *L'Information Grammaticale*, *136*(1), 25–32. [CrossRef]

Lorenzi Bailly, N., & Moïse, C. (Eds.). (2021). *La haine en discours*. Le Bord de L'eau.

Moeschler, J., & De Spengler, N. (1981). Quand même: De la concession à la réfutation. *Cahiers de linguistique française*, *2*, 93–112. Available online: https://www.unige.ch/clf/fichiers/pdf/07-Moeschler_nclf2.pdf (accessed on 6 March 2025).

Moeschler, J., & De Spengler, N. (1982). La concession ou la réfutation interdite. *Cahiers de linguistique française*, *4*, 7–36. Available online: https://www.unige.ch/clf/fichiers/pdf/02-Moeschler_nclf4.pdf (accessed on 6 March 2025).

Moïse, C. (2020). Pour (re)venir à une sociolinguistique du sujet et de la subjectivité. In K. D. Léonard, & V. Rose-Marie (Eds.), *Appropriation des langues et subjectivité. Mélanges offerts à Jean-Marie Prieur par ses collègues et amis* (pp. 59–70). L'Harmattan.

Moïse, C., Meunier, E., & Romain, C. (2019). *La violence verbale dans l'espace de travail: Analyses et solutions* (2nd ed.). Bréal.

Moïse, C., & Oprea, A. (2015). Présentation. Politesse et violence verbale détournée. *Semen. Revue de Sémio-Linguistique des Textes et Discours*, *40*, 2–12. [CrossRef]

Moreau Raguenes, R. (2021). *La construction du sujet agentif sur Parents toxiques: Analyse discursive d'un (re)maniement de soi par le témoignage de maltraitance parentale* [Master's dissertation, Université Grenoble Alpes]. Available online: https://dumas.ccsd.cnrs.fr/dumas-03610362 (accessed on 6 March 2025).

Moreau Raguenes, R. (2022). (Dé)montrer la maltraitance parentale sur Instagram: Étude argumentative du récit extime catégorisant. *Cahiers de Narratologie. Analyse et Théorie Narratives*, *42*, 1–20. [CrossRef]

Moreau Raguenes, R. (2024a). L'être comme fautif. Actes de condamnation, altérité et rapport de places dans l'interaction parent-enfant. *SHS Web of Conferences*, *191*, 01009. [CrossRef]

Moreau Raguenes, R. (2024b). Saisir la maltraitance parentale. Une approche discursive. In S. Wharton, S. Vernet, & M. Gasquet-Cyrus (Eds.), *La sociolinguistique, à quoi ça sert? Sens, impact, professionnalisation* (pp. 163–177). Presses Universitaires de Provence. Available online: https://hal.science/hal-04831025 (accessed on 6 March 2025).

Morel, M.-A. (1996). *La concession en français*. Ophrys.

Neuburger, R. (2008). *L'art de culpabiliser*. Payot.

Rosier, L. (1999). *Le discours rapporté: Histoire, théories, pratiques*. Duculot.

Rosier, L. (2008). *Le discours rapporté en français*. Ophrys.

Rosier, L. (2009). *Petit traité de l'insulte*. Esprit libre.

Searle, J. R. (1982). *Sens et expression: Études de théorie des actes de langage* (J. Proust, Trans.). Les Éditions de Minuit.

Taylor, C. (2015). Beyond sarcasm: The metalanguage and structures of mock politeness. *Journal of Pragmatics*, *87*, 127–141. [CrossRef]

Van Hooland, M. (Ed.). (2005). *Psychosociolinguistique: Les facteurs psychologiques dans les interactions verbales*. L'Harmattan.

Van Hooland, M. (2008). L'enfant et sa stratégie discursive d'adaptation en situation de maltraitance familiale, approche psychosociolinguistique des interactions verbales maltraitantes. In C. Moïse, N. Auger, B. Fracchiolla, & C. Romain (Eds.), *La violence verbale. Tome 2. Des perspectives historiques aux expériences éducatives* (pp. 215–236). L'Harmattan.

Vincent, D. (2013). L'agression verbale comme mode d'acquisition d'un capital symbolique. In B. Fracchiolla, C. Moïse, C. Romain, & N. Auger (Eds.), *Violences verbales. Analyses, enjeux et perspectives* (pp. 37–54). Presses Universitaires de Rennes.

Vincent, D., Laforest, M., & Turbide, O. (2008). « Pour un modèle fonctionnel d'analyse du discours d'opposition: La trash radio ». In C. Moïse, N. Auger, B. Fracchiolla, & C. Schulz-Romain (Eds.), *La violence verbale*. L'Harmattan.

languages

MDPI

Article

"She'll Never Be a Man" A Corpus-Based Forensic Linguistic Analysis of Misgendering Discrimination on X

Lucia Sevilla Requena

Department of English Philology, Universidad de Alicante, Carr. de San Vicente del Raspeig, 03690 Alicante, Spain; lsr30@alu.ua.es

Abstract: Misgendering is a form of microaggression that reinforces gender binarism and involves the use of incorrect pronouns, names or gendered language when referring to a transgender and gender non-conforming (TGNC) individual. Despite growing awareness, it remains a persistent form of discrimination, and it is crucial not only to understand and address misgendering but also to analyse its impact within online discourse towards the TGNC community. The present study examines misgendering directed at the TGNC community present on platform X. To achieve this, a representative sample of 400 tweets targeting two TGNC individuals is compiled, applying an annotation scheme to manually classify the polarity of each tweet and instances of misgendering, and then comparing the manual annotations with those of an automatic sentiment detection system. The analysis focuses on the context and frequency of intentional misgendering, using word lists to examine the data. The results confirm that misgendering perpetuates discrimination, tends to co-occur with other forms of aggression, and is not effectively identified by automatic sentiment detection systems. Finally, the study highlights the need for improved automatic detection systems to better identify and address misgendering in online discourse and provides potentially useful tools for future research.

Keywords: misgendering; forensic linguistics; microaggressions; TGNC community; corpus annotation; natural language processing (NLP)

Citation: Sevilla Requena, Lucia. 2024. "She'll Never Be a Man" A Corpus-Based Forensic Linguistic Analysis of Misgendering Discrimination on X. *Languages* 9: 291. https://doi.org/10.3390/languages9090291

Academic Editors: Julien Longhi and Nadia Makouar

Received: 7 July 2024
Revised: 23 August 2024
Accepted: 24 August 2024
Published: 30 August 2024

1. Introduction

Gender identity is a fundamental aspect of the human being that reflects the inner sense of who we are. Recognition and respect for gender identity are essential to build an inclusive society, where each individual can live following their true identity without fear of reprisal or discrimination.

Transgender and gender non-conforming (hereafter TGNC) individuals represent a complex and diverse population whose gender identities defy conventional societal norms (Argyriou 2021, p. 71). Rooted in a deep sense of self-awareness and authenticity, these individuals experience a profound misalignment between their inner sense of gender and the sex they were assigned at birth (American Psychological Association 2015).

Furthermore, within the TGNC community, a rich tapestry of experiences and identities exists. Some individuals choose to undergo physical transitions, such as hormone therapy or surgical procedures, to align their bodies with their gender identity. In contrast, others opt for social transitions, changing their name, pronouns and outward presentation to reflect their true selves (Argyriou 2021, p. 71).

Despite this diversity within the TGNC community, individuals often face challenges in how they are perceived and, consequently, addressed by others. An example of this is the linguistic phenomenon of 'misgendering', a concept that has attracted increasing attention from scholars and the media, and that refers to the act of using incorrect pronouns or gendered language when referring to TGNC individuals (McLemore 2016). According to Argyriou (2021), "everyday life of TGNC people is filled with examples of invalidations of

the kind, as misgendering is generalised and persistent" (p. 72). This can manifest in overt forms of disrespect, such as deliberate misidentification or verbal harassment, as well as more subtle forms of invalidation, such as dismissing a person's gender identity or ignoring their preferred pronouns (Argyriou 2021, p. 73), who can sometimes appear together.

The consequences of these forms of disrespect, either overt or subtle, when performed repeatedly during a long period, are profound and have severe implications of cultivating a hostile atmosphere that can adversely affect the mental and emotional well-being of the targeted individuals. Moreover, when these acts become the norm, they reinforce harmful stereotypes and perpetuate systemic discrimination against TGNC people. They can have, as a final consequence, the neglect of access to mainstream society, which is the ultimate goal of hate speech (Guillén-Nieto 2023).

Hence, it is necessary to gain a deep understanding of the complexities surrounding misgendering, as it involves delving into the context of several different TGNC populations' experiences. Each TGNC person's journey is unique and shaped by various factors, including culture, background, and personal beliefs. Therefore, the attempt to understand and identify instances of misgendering faces a series of difficulties, which leads to the creation of the present study.

The primary aim of this study is to disseminate the concept of intentional misgendering, as a manifestation of discrimination expressed through language, to facilitate its subsequent detection on social networks, specifically platform X[1]. To accomplish this goal, a series of more specific objectives will be pursued:

- Create a dataset of tweets targeting TGCN individuals from platform X, applying a set of criteria to ensure relevance and accuracy.
- Implement an annotation scheme to classify each tweet's polarity and evaluate the consistency of the manual annotation between two annotators.
- Examine the context and frequency of those tweets that include intentional misgendering, analysing wordlists of the instances.
- Evaluate the effectiveness of an automatic sentiment detection system by comparing its performance to manual annotations.
- Provide recommendations for improving automatic detection systems and addressing intentional misgendering in online discourse effectively.

By pursuing these interconnected objectives, this study aims to expand the understanding of the linguistic phenomenon of intentional misgendering, ultimately contributing to the creation of safer and more inclusive online environments for all users.

2. Theoretical Background

Language, as a fundamental communication tool, can be weaponised to inflict harm, with words carrying both damaging and legally significant consequences. Forensic linguistics is key in examining how such harmful language can serve as evidence in legal contexts (Guillén-Nieto 2022, p. 1). In this context, this section offers an in-depth review of essential concepts central to understanding harmful language, starting with the broader framework of harassment, progressing through microaggressions, and ultimately focusing on misgendering, which is the core of this study. The purpose is to clarify how these forms of derogatory language interrelate and provide a more comprehensive understanding of their objectives and their legal and social implications.

2.1. The Concept of Harassment

Harassment is a highly complex and multifaceted phenomenon that encompasses a wide range of offensive behaviours designed to undermine an individual's dignity. These behaviours, which often consist of hostile and unethical forms of communication, are directed by one or multiple perpetrators victimising a particular target for an extended period (Leymann 1990). According to Guillén-Nieto (2022, p. 7), it involves a series of acts directed towards the destruction or diminution of the fundamental rights of the affected individual. The objective behind such actions is typically malicious, mainly due

to the desire of perpetrators to achieve certain aims or goals, which may vary depending on the context. For instance, the case of gender-based harassment faced by the TGNC community contributes to the central goal of hate speech, which is to deny them equal access and exclude them from the rest of society. This type of harassment, considered a macro-directive, is carried out through the execution of a series of micro-acts of aggression, each of which indirectly contributes to the achievement of these super-goals (Guillén-Nieto 2022, p. 8). Hence, the series of smaller aggressive actions conforming to harassment receive the name of 'microaggressions'.

2.2. The Concept of Microaggressions

In 1969, Dr Chester Pierce introduced the term 'offensive mechanisms' to describe the subtle but pervasive ways in which black people were marginalised in the United States. Pierce noted, "To be black in the United States today means to be socially minimised. Every day, black people face 'offensive mechanisms' designed to isolate, diminish, and confine them to a lesser status. The relentless message they receive is that they are unimportant and irrelevant" (Pierce 1970, p. 303). In a 1970 essay titled "Offensive Mechanisms", Dr Pierce further developed this concept, coining the term 'microaggressions' to refer to these understated yet impactful actions.

Additionally, psychologist Sue (2010) defined microaggressions as "brief, everyday interactions that convey negative or derogatory messages to people because of their identity group" (p. 36). Over time, the term has come to include not just racial bias but also insults and behaviours targeting other marginalised groups, such as ethnic minorities, gender minorities and people with disabilities (Sue 2010; Paludi 2012).

When targeting the TGNC community, the subtlety of microaggressions does not diminish their impact since it can lead to significant repercussions, including undermining a person's identity and invalidating their existence. Therefore, the TGNC individuals are faced with a profound dissonance between their gender identity and the sex they were assigned at birth, which makes them vulnerable to misgendering—a form of microaggression that involves being addressed or named in a manner that is inconsistent with their gender identity (Argyriou 2021, p. 72).

2.3. The Concept of Misgendering

The American Psychological Association, in its "Guidelines for Psychological Practice with Transgender and Gender Nonconforming People," defines 'misgendering' as "Using pronouns or other words that label a person's gender incorrectly. This is often a painful experience for people including trans and gender nonconforming people, especially when done by someone aware of their gender identity." (American Psychological Association 2015). This phenomenon serves as a potent reminder of the broader societal issues surrounding gender identity and acceptance and is crucial to combat this harmful practice with education, research and awareness. Thus, this study highlights misgendering as a central issue.

3. State of the Art

The literature review presents a detailed examination of research on 'microaggressions' and 'misgendering' from a linguistic perspective, to compile and present prior studies on the analysis and annotation of gender microaggressions. This review is divided into two main sections: descriptive linguistics and corpus linguistics approaches to gender microaggressions, focusing particularly on intentional misgendering.

3.1. Descriptive Linguistics Approach to Gender Microaggressions

Microaggressions have traditionally been examined in the context of racial and ethnic discrimination (Chang and Chung 2015, p. 220). Sue and Capodilupo (2008) were the first to identify parallels between racial and gender microaggressions, suggesting that the mechanisms underlying these biases might share commonalities.

In addition, Solórzano et al. (2000) was the first to coin the term 'gender microaggressions', but there has been limited empirical research to substantiate the concept. However, only in recent years has there been a significant expansion in the scope of research, encompassing microaggressions directed toward the LGBTQIA+ community. This broader focus reflects an increasing awareness of how microaggressions manifest, highlighting the need for further empirical investigations to understand and address these subtle but impactful forms of discrimination.

According to Nadal et al. (2016), research on microaggressions within the lesbian, gay and bisexual (LGB) communities has increased, yet studies addressing TGNC individuals remain sparse. This shortfall is partly due to the common conflation of gender identity with sexual orientation in broader LGBTQIA+ studies, which obscures the unique experiences of transgender people who face different forms of discrimination and marginalisation compared to their LGB counterparts (Fassinger and Arseneau 2007; McCarthy 2003). Despite inclusive intentions, combining these identities can inadvertently perpetuate the marginalisation of TGNC voices.

Sue et al. (2008) developed a classification system for gender-based microaggressions, which was later updated by Nadal et al. (2010). Building upon this work, Nadal et al. (2012) conducted a qualitative study that explored the nuanced interpersonal and systemic microaggressions faced by TGNC individuals. Participants from diverse backgrounds were recruited through local LGBTQIA+ organisations to form two focus groups. The study aimed to validate the existing taxonomy of microaggressions towards transgender people through directed content analysis, revealing twelve themes specific to TGNC individuals, thus expanding the understanding of their experiences.

Some of the themes identified included "the use of transphobic or incorrectly gendered terminology", which involves derogatory terms or incorrect pronouns; "the assumption of a universal transgender experience", which stereotypes transgender individuals; or "the exoticisation", where transgender individuals are fetishised.

Nadal et al. (2016) observed that one of the most prevalent forms of microaggression encountered by TGNC individuals is the failure to recognise or consistently use their preferred pronoun, "particularly after someone has been corrected or informed of a genderqueer person's preferences" (p. 13). Given this, the present study will concentrate on the first category identified in the taxonomy—namely, the use of transphobic or incorrectly gendered terminology—focusing on the phenomenon of 'misgendering'. Misgendering involves the erroneous attribution of gender to an individual. This misattribution can materialise through the use of pronouns, titles or descriptions (McNamarah 2021). Moreover, the repercussions of such a practice can be highly damaging and deeply detrimental for those who experience it, as it undermines their sense of identity and belonging.

Particularly, this study focuses on 'intentional misgendering', which is characterised by a conscious and deliberate decision to disregard an individual's preferred gendered language or titles. Unlike unintentional misgendering, which may stem from ignorance or oversight, intentional misgendering involves a deliberate choice to ignore or reject the correct gender designation of the person being addressed (McNamarah 2021, p. 2261).

In the literature on transgender identities and microaggressions, the phenomenon of misgendering has garnered significant attention. A recent study by Edmonds and Pino (2023) provided an in-depth analysis of intentional misgendering and its effects on power dynamics, identity construction and societal norms. Their research found that intentional misgendering is used to undermine transgender individuals' identities, revealing how such acts are strategically employed to express negative attitudes and reinforce cisgenderism views. Trans individuals often respond by framing misgendering as morally reprehensible, challenging the discriminatory behaviour of the offenders.

Another study by Thál and Elmerot (2022) delved into the misgendering of transgender individuals in the Czech language, identifying linguistic infra-humanisation and dehumanisation as key components. The study highlighted the importance of recognising and addressing linguistic hostility towards transgender individuals through the analysis of

diverse textual sources. This research underscores the need for sensitivity in language use to avoid reinforcing harmful practices and calls for greater awareness and inclusivity in communication to support TGNC individuals better.

However, despite the recent interest in this topic, there is still a gap in the literature, and this study seeks to fill it by exploring the phenomenon of intentional misgendering in English and collecting and annotating a corpus sample of tweets from platform X directed at TGNC individuals.

3.2. Corpus Linguistics Approach to Microaggressions Annotation

Within the framework of corpus linguistics, a significant gap is also identified in the literature regarding misgendering as a linguistic phenomenon. While misgendering, as a form of microaggression, has been subject to theoretical study for years, there is a notable absence of computational analyses on this topic. This gap represents a great opportunity to further explore and expand linguistic annotation efforts on the phenomenon of misgendering, which can serve the purpose of enriching language understanding and improving the inclusivity and accuracy of NLP technologies in various contexts.

Relevant work on gender bias annotation is the one by Havens et al. (2022), which addresses the challenge of mitigating gender bias in NLP systems by developing a taxonomy of gendered language and applying it to create annotated datasets. The taxonomy categorises labels into three main categories: Person Name, Linguistic, and Contextual, with sub-labels defined by archival documentation. The study underscores the significance of interdisciplinary collaboration and clear metrics in identifying gender bias and aims to guide NLP systems towards more inclusive representations of gender.

Assimakopoulos et al. (2020) conducted a study that introduces a hierarchical annotation scheme to discern discriminatory comments within online discussions. The scheme goes beyond the binary classification of hate speech versus non-hate speech, categorising comments based on their attitude towards target minority groups and detailing how negative attitudes are articulated. The study found that this multi-level scheme improved inter-annotator agreement, highlighting the need for refined annotation guidelines and comprehensive training for annotators to identify negative discourse strategies better.

These studies collectively highlight the importance of annotation in corpus-based research, particularly in identifying linguistic features and extracting insights from textual data. Annotation is crucial in advancing fields such as NLP by facilitating the systematic categorisation of elements within textual data. In the context of gender-based discrimination, effective annotation frameworks are essential for accurately capturing the complexity and impact of intentional misgendering.

The present research aims to fill the corpus collection and annotation gap related to misgendering. Developing robust annotation frameworks can enhance the detection of intentional misgendering, contributing to a more inclusive and respectful environment for TGNC individuals. This can help to address the root causes of misgendering and promote greater awareness of the importance of using correct and respectful language through focused corpus collection and annotation.

4. Research Questions

The present study formulates relevant research questions that will allow for a deeper investigation of the linguistic phenomenon presented in this study and the aspects that encompass it.

The questions proposed are as follows:

- RQ1: Does intentional misgendering as a form of microaggression perpetuate discrimination towards the TGNC community?
- RQ2: Does intentional misgendering typically co-occur with other forms of aggression or discriminatory language?
- RQ3: Is there a significant relationship between the presence of misgendering in tweets and their sentiment polarity?

- RQ4: Can automatic sentiment detection systems effectively identify tweets containing misgendering and expressing hatred towards transgender individuals, or is there a gap in their ability to detect this type of message?

These research questions will guide the compilation of messages extracted from platform X, as well as the evaluation of manual annotation and automatic polarity detection systems on misgendering directed at TGNC.

5. Methodology

5.1. Corpus Compilation

The corpus sample consists of a collection of tweets in English extracted from platform X, specifically focusing on intentional misgendering within the context of online discourse. This corpus sample aims to analyse and identify instances of intentional misgendering targeting two specific TGNC public figures. In addition, the text type selected for creating the corpus sample consists of tweets that mention two TGNC individuals who are public figures. This choice of text type is motivated by the prevalence of X as a platform for public discourse and its potential for capturing informal communication that reflects broader societal attitudes. Lastly, the selection of English as the language for the corpus sample is due to its widespread use on social media platforms and its relevance related to this Master.

5.1.1. Data Selection Criteria

To construct the corpus for studying intentional misgendering targeting the TGNC community, specific selection criteria have been established. In the words of Biber (1993), "a corpus must be 'representative' in order to be appropriately used as the basis for generalisations concerning a language as a whole" (p. 243). For this reason, the criteria selected to ensure the representativeness of the phenomenon in the corpus sample are the publication dates, size, authors and selection of the individuals targeted by misgendering.

Regarding the publication dates of the texts, tweets have been selected within a time range spanning from 1 January 2023, to the date of the compilation (15 April 2024). The choice of this period is based on significant changes in the "Abuse and Harassment Policies" of the social network X, particularly the removal and subsequent reintroduction of rules against misgendering between 2023 and 2024 (X 2024).

As for the sample size selection, it has been divided to represent both individuals equally. In total, 400 tweets were selected, with 200 tweets directed at each individual, creating a balanced framework for comparison and analysis, and ensuring representativeness. This sample size has been deliberately chosen to support the study's qualitative approach, prioritising a deep, nuanced exploration of microaggressions to understand the complex nature of the phenomenon.

Considering the author of the corpus, it is not possible to establish specific criteria for the authors of the tweets since each tweet originates from different individuals with varied backgrounds, styles and motivations. Given this diversity, there is not a uniform set of criteria that can accurately categorise or classify them.

To further outline the corpus sample criteria, it is essential to explain the selection of the individuals targeted by misgendering. The choice is based on their public prominence and the controversies surrounding their transitions to ensure a certain number of tweets are directed at them.

In addition, it is important to remark that as Burghardt (2015) states, "redistributing Twitter content outside the Twitter platform is prohibited. In practice, this means it is not possible to precompile Tweet corpora and to share them in a way they are readily accessible for academic research." (p. 78). This means that to carry out the corpus sample compilation, the individuals need to face an anonymisation process where their names are changed to hide their identity. To carry out the corpus sample compilation, the individuals must face an anonymisation process where their names are changed to hide their identity. However, misgendering is inherently contingent upon the identity, context and cultural setting within which it occurs, as noted by Hochdorn et al. (2016). For that reason, it is necessary to briefly

contextualise the situation of the individuals to understand the reasoning behind the choice for this study:

- "Individual 1" is a transgender man famous for his work as an actor before his transition. His pronoun set is he/they.
- "Individual 2" is a transgender woman known for documenting her transition process on social media. Her pronoun set is she/they.

After establishing the criteria for the corpus sample, the next step in this stage involves extracting tweets that meet these criteria.

5.1.2. Data Extraction

The first approach to extracting tweet data involves using the Twitter Search API to access tweets from the 'Top' category. For this method, tweets containing the keywords "Individual 1" or "Individual 2" from 1 January 2023 to 15 April 2024 are retrieved. To enhance the data collection, a second approach using web scraping methods is also employed. This complementary method involves creating a query search on X and scraping the web code to extract each of the tweets one by one.

5.1.3. Data Pre-Processing

The raw data obtained from X underwent a pre-processing stage to refine it into a more usable format. Given that raw data often include numerous irrelevant tweets, a manual effort was made to simplify the dataset, thereby minimising the annotation of tweets not fitting the established criteria. To achieve this, the following filtering mechanisms devised to selectively extract tweets were employed:

- Keyword presence: Tweets explicitly mention the keyword "Individual 1" or "Individual 2".
- Duplicate tweets: Excluding tweets that are duplicates.
- URL/image tweets: Filtering out tweets solely consisting of URLs/images.
- Language criterion: Eliminating tweets composed in languages other than English or those containing a mix of English and non-English content.
- Minimum length: Disregarding tweets with a character count below five.
- User mentions: Removing tweets primarily comprised of user[2] mentions.

Thus, implementing this filtering process ensures the efficiency of subsequent annotation efforts and the dataset's quality and integrity.

5.1.4. Dataset Statistics

The corpus sample is composed of 400 tweets, with an equal distribution of 200 tweets referring to Individual 1 and 200 referencing Individual 2. The complete corpus contains a total of 14,492 tokens and 12,284 unique types. However, to ensure a comprehensive examination of misgendering patterns, it was divided into two subsets or sub-corpus based on whether the tweets mention Individual 1 or 2. This division facilitates a targeted analysis of misgendering, recognising that its patterns may differ when directed at a trans man versus a trans woman. The first subset, comprising tweets referencing Individual 1, includes 7458 tokens and 6276 unique types, while the second subset, centred on Individual 2, contains 7034 tokens and 6008 unique types.

In addition, a wordlist frequency was generated for both subsets using Sketch Engine[3]. This platform provides a detailed breakdown of the common words and phrases from the corpus sample using the corpus enTenTen21[4] as a reference (Suchomel 2020). The following wordlists allow a more precise analysis of the language surrounding misgendering (see Tables 1 and 2).

Table 1. Wordlist of sub-corpus 1 from Sketch Engine.

No.	Lemma	Frequency	No.	Lemma	Frequency
1	Trans	26	11	Delusion	5
2	Lesbian	10	12	Tittie	5
3	Tit	9	13	Cisgender	3
4	Transitioned	8	14	Mutilate	3
5	Transphobe	7	15	Psychotic	3
6	Mastectomy	6	16	Deadnaming	2
7	Topless	5	17	Self-hatred	2
8	Cis	5	18	Weirdo	2
9	Slur	5	19	Transman	2
10	Trans	5	20	Objectify	2

Table 2. Wordlist of sub-corpus 2 from Sketch Engine.

No.	Lemma	Frequency	No.	Lemma	Frequency
1	Trans	18	11	Fiasco	3
2	Pretend	10	12	Womanhood	3
3	Gay	8	13	Manly	3
4	Dude	6	14	Backlash	3
5	Girlhood	4	15	Clown	3
6	Mock	4	16	Transgender	3
7	Pronoun	4	17	Influencer	3
8	Marginalise	4	18	Vomit-inducing	2
9	Leftist	4	19	Transvestite	2
10	Mockery	3	20	Transphobic	2

Overall, the combination of the corpus sample subsets and the wordlist frequency analysis provides a comprehensive foundation for understanding the complexities of misgendering in online discourse, laying the groundwork for further exploration and research in this critical area. These keyword tables will be used in the analysis of the results alongside the annotation data to examine the relationship between these terms and their associated sentiments, as well as to evaluate their connection to misgendering.

5.2. Corpus Annotation

After the data selection criteria, extraction and pre-processing procedures are completed, the second stage involves manually annotating the corpus sample. The upcoming sections detail the design of the annotation scheme and the steps taken in the annotation process to create a reliable and accurate corpus sample.

5.2.1. Annotation Scheme

The annotation scheme used in this corpus sample is designed to capture the sentiment expressed in the tweets while maintaining a consistent and reliable classification system. This scheme consists of three main polarity groups: neutral, negative or positive, which allows for a simple but comprehensive assessment of sentiment towards TGNC people in the extracted tweets. This methodology is outlined in the SemEval[5] task by Rosenthal et al. (2015). Moreover, the simplicity and clarity of the selected annotation scheme contribute to reducing the ambiguity that could arise in such a context-dependent phenomenon as intentional misgendering. Some illustrative examples of tweets and their corresponding annotations are presented in Tables 3 and 4. Per X's anonymisation policies, the tweets have been paraphrased to ensure compliance.

Table 3. Annotation of sentiment polarity in sub-corpus 1.

Polarity	Confidence	Tweets
Positive	1	I had an unusual dream about Individual 1 and now I think I might have feelings for him, lol.
Negative	1	She'll never be a man, no matter what she does. Even though Individual 1 has invested a lot in trying to transition, people will not perceive her as part of our group.
Neutral	1	Film 1-directed by director 1, starring actor 1 and executive-produced by Individual 1-centers around a teenage girl navigating a competition.

Table 4. Annotation of sentiment polarity in sub-corpus 2.

Polarity	Confidence	Tweets
Positive	1	if Individual 2 did single-handedly disrupt the entire product 1 industry, she must be one of the most influential women in the world.
Negative	1	Individual 2: "I do not think God made an error with me." He seems to be trying to sway people but inadvertently speaks the truth. Indeed, God didn't make an error with him—he was made a man.
Neutral	1	Individual 2 was the winner of the first Woman of the Year award by brand 1.

5.2.2. Annotation Process

The annotation process involves several critical steps to ensure consistency and reliability. Initially, the tweets were assigned to two annotators with advanced Linguistics and English knowledge. Annotators were tasked with determining the overall polarity of each tweet based on their sentiment towards the targeted individuals, classifying them into one of three categories: neutral, negative or positive. Each annotator works independently to minimise bias and subjectivity during this classification stage.

Following the independent annotation, a reconciliation phase took place. During this phase, any discrepancies or disagreements in the sentiment classification are discussed and resolved through the final decision of a referee. This step is crucial to maintain consistency and ensure the annotations accurately represent the sentiment in the tweets.

Finally, after the reconciliation phase, the annotated corpus sample was compiled, along with a detailed record of the annotation decisions, including any resolved disagreements. This final corpus served as the foundation for further data analysis, with the annotations providing a structured framework for exploring sentiment and misgendering trends.

5.2.3. Annotation Guidelines

To ensure consistency and accuracy in the annotation process, specific guidelines have been established for annotators to ensure that the dataset is reliable and consistent. These guidelines are designed to help annotators focus on the annotation's main objective, which is to determine the polarity of the tweets.

The first thing to consider when annotating sentiment, is whether the language used in the text respects or misrepresents the gender identity of the individuals being discussed. This involves examining the tone, word choice and contextual elements that indicate whether the sentiment is positive, neutral or negative. After determining the sentiment polarity, annotators should use a confidence scale from 0 to 1 to indicate the level of certainty

for each annotation. This confidence measure helps to identify annotations that might require further review or discussion.

Furthermore, annotators are required to identify misgendering since it is the main research object in the present study. This procedure ensures all annotators can correctly recognise misgendering, providing a uniform understanding and approach to the annotation task.

5.2.4. Inter-Annotator Agreement

When annotating corpora involving multiple human annotators, it is critical to ensure consistency and reliability (Artstein and Poesio 2008, p. 556). This requirement is universal across all annotation types but is particularly crucial for corpus containing ambiguous or subjective factors such as intentional misgendering. This is because, with complex subjects, annotators might interpret annotation guidelines differently, leading to variation in their annotations. The inconsistency can defeat the purpose of the annotated corpus, as it can impede machine learning algorithms from extracting useful patterns for predictions. Thus, assessing the reliability of the annotation process is essential, using specific metrics to ensure high-quality outcomes (Moreno-Ortiz and García-Gámez 2022, p. 192).

To measure the consistency of the annotations, Cohen's Kappa statistic (Cohen 1960, p. 42) was used. This metric accounts for the likelihood of agreement occurring by chance, providing a more robust measure of inter-annotator reliability than simple percentage agreement. In this instance, the kappa statistic was calculated on a subset of 100 tweets, out of 400, annotated by the two annotators. The resulting kappa value was 0.8168, indicating a high level of agreement between the annotators. This high kappa value suggests that the annotation process is consistent, providing confidence in the quality of the annotated data and supporting the reliability of subsequent analyses and machine learning applications.

6. Results and Discussion

This section presents the results obtained from the compilation and annotation of the corpus sample, providing a comprehensive analysis of the data to elucidate the nature and implications of intentional misgendering. The findings are quantified and exemplified to offer a detailed overview of the corpus.

The final annotation results provide valuable insights into two important aspects of the corpus: sentiment polarity and misgendering patterns. Sentiment polarity refers to whether tweets express a positive, negative or neutral sentiment towards the TGNC individuals. In contrast, intentional misgendering is annotated referring to the use or misuse of gender pronouns and gendered language to describe the two individuals. To start, Table 5 provides an overview of the tweets classified by the polarity of sentiment by the two annotators. The data suggest a strong negative bias in the sentiment of the overall corpus sample.

Table 5. Sentiment polarity of the corpus sample.

	Positive	**Negative**	**Neutral**	**Total**
Individual 1	57	126	17	200
Individual 2	26	153	21	200
Total	83	279	38	400

The aggregated data across both manual annotators reveal that negative tweets significantly outnumber positive and neutral ones, with 279 negative tweets compared to 83 positive and 38 neutral tweets. Notably, more than half of the tweets exhibit negative sentiments towards both individuals, underscoring the prevalence of derogatory content targeting TGNC individuals. These initial findings call for further research to explore the relationship between these sentiments and intentional misgendering.

Moreover, a more detailed analysis of misgendering was deemed necessary to determine whether this linguistic phenomenon differs based on the gender identity of the

targeted individual. This approach aimed to identify and clarify any variations in how misgendering manifests when directed toward a trans man (Individual 1) versus a trans woman (Individual 2). Consequently, a sub-annotation process was conducted to categorise instances of intentional misgendering.

For this sub-annotation process, the data were analysed using a predefined set of classifications by McNamarah (2021) designed to capture the different forms of misgendering. By using these classifications, annotators could systematically identify and categorise instances of misgendering directed at both Individuals 1 and 2. This process aimed to provide a comprehensive understanding of misgendering in this context.

1. The tweets that exhibit "mislabelling", which involves using incorrect gendered terms or categories that do not align with the individual's gender identity, are annotated as MISLABEL.
2. The tweets that exhibit "mispronouning", which entails using incorrect pronouns when addressing or referring to the individual, disregarding their gender identity, are annotated as MISPRONOUN.
3. The tweets that use the correct pronouns or gendered language when referring to the individuals, aligning with their stated gender identity, are annotated as CORRECT GENDER.
4. The tweets that do not directly address the individual's gender identity and do not specify any gender-specific treatment are annotated as NO.

Other forms of misgendering, such as deadnaming, were excluded because the primary criterion for tweet inclusion was the use of the individuals' chosen names post-transition. Consequently, ungendering and unpronouning were also excluded since both individuals have "they" as part of their pronoun set.

6.1. Analysis of the Results

The subsequent set of Tables 6 and 7 aims to provide a more nuanced understanding of the polarity of sentiment in tweets, distinguishing between those containing instances of misgendering and those without. This division is achieved by detailing the specific type of misgendering used in each tweet.

Table 6. Annotation of misgendering in sub-corpus 1.

Misgendering	Positive	Negative	Neutral	Total
NO	18	20	9	47
CORRECT_GENDER	40	1	6	47
MISLABEL	0	27	1	28
MISPRONOUN	0	78	1	79
Total	57	126	17	200

For tweets where pronouns or gendered words targeting Individual 1 are absent (NO), the sentiment distribution was 20 negative tweets compared to 18 positive and 9 neutral. Among those that correctly used male pronouns and gender language (CORRECT GENDER), the distribution varied slightly, with 40 tweets that exhibited positive sentiment, only 1 tweet annotated as negative and 6 considered neutral.

For tweets containing mispronouning (MISPRONOUN), where pronouns like "she", "her" or both together are used to refer to Individual 1, the data indicate a strong negative bias. Of the tweets with mispronouning, 78 are annotated as negative, with only 1 neutral. Regarding mislabelling (MISLABEL), the tweets were also predominantly negative, with 27 negative, 1 neutral and no positive tweets, and the words employed were mostly "woman" and "girl".

Overall, the data suggest a correlation between misgendering and negative sentiment, with 126 tweets annotated as negatives, 78 presenting mispronouning and 27 with mis-

labelling. The high frequency of negative sentiment among tweets with misgendering indicates that such language is often used in a derogatory or hostile context when targeting Individual 1. Furthermore, the consistent pattern of misgendering with negative sentiment underscores the importance of further exploration to understand the underlying causes and the broader implications for TGNC individuals in social media discourse.

In addition, to assess the statistical validity of the data shown in Table 6, a Chi-Square test was conducted, resulting in a p-value < 0.05. This low p-value indicates that the differences observed in the data are not likely due to random chance, making the results statistically significant.

Table 7. Annotation of misgendering in sub-corpus 2.

Misgendering	Positive	Negative	Neutral	Total
NO	17	49	18	84
CORRECT_GENDER	9	1	2	12
MISLABEL	0	33	1	34
MISPRONOUN	0	70	0	70
Total	26	153	21	200

Additionally, when examining tweets targeting Individual 2, a similar pattern emerges concerning the relationship between misgendering and sentiment polarity. Tweets where pronouns or gendered terms are absent (NO) exhibit a negative sentiment bias, with 49 non-negative, 17 positive, and 18 neutral tweets.

In contrast, tweets that use correct female pronouns or gendered language (CORRECT GENDER) show a more positive distribution, with nine positive, one negative and two neutral tweets. However, compared to tweets directed at Individual 1, the number of tweets using the correct gender for Individual 2 is much lower. This could suggest a difference in how people refer to trans women compared to trans men, which may occur because of deep-rooted social perceptions and stereotypes that influence the way people use language to describe transgender individuals. Additionally, this discrepancy in using correct gender pronouns might also indicate that there is less public recognition or visibility for trans women compared to trans men. Trans women experience both transphobia and misogyny, creating a double burden of discrimination, which can eventually lead to a reduced disposition among the public to use correct gender pronouns, either through ignorance or deliberate disregard.

For tweets containing mislabelling (MISLABEL), where terms like “man” or “dude” are used to describe Individual 2, there is a significant bias towards negative sentiment. From these tweets, 33 are annotated as negative, with only 1 classified as neutral and none as positive. This indicates that mislabelling often conveys a derogatory tone. Similarly, tweets with mispronouning (MISPRONOUN), which use male pronouns like “he”, “his” or “him” to refer to Individual 2, exhibit an overwhelmingly negative sentiment. All 70 tweets with mispronouning are annotated as negative, without any positive or neutral classification.

Again, to assess the statistical validity of the data presented for Individual 2 in Table 7, a Chi-Square test was conducted, resulting in a p-value < 0.05. This indicates that the differences observed in the data are statistically significant and unlikely to have occurred by random chance.

To summarise, the data reveal a clear correlation between misgendering and negative sentiment. Among the total 200 tweets analysed, 153 are negative, with a significant portion (70) involving mispronouning and 33 containing mislabelling. The data indicate that misgendering language is often associated with derogatory or hostile contexts, emphasising the need to explore the underlying reasons behind this pattern and its implications for Individual 2 and other TGNC individuals in social media discourse.

6.2. Analysis of the Automatic Annotation

In this section, a comprehensive examination of the annotation results is conducted to uncover annotation issues. To achieve this, the section is subdivided into two parts based on the forms of annotation performed for this corpus sample: manual and automatic. An analysis of these annotations and an improvement of their guidelines is explored for further research. Furthermore, the section also covers an in-depth analysis of sentiment polarity and misgender trends observed in the corpus sample.

6.2.1. Automatic Annotation Issues

In the following section, the Python library flairNLP v0.13.1 (Akbik et al. 2019) is used to conduct sentiment analysis using deep learning models. Flair is a user-friendly NLP framework that offers pre-trained models for various tasks, including sentiment analysis. Despite its robustness, discrepancies between automated systems and human experts are common. These inconsistencies can be due to several factors, and understanding their root causes is crucial for enhancing the reliability of automated sentiment analysis (Birjali et al. 2021; Kozareva et al. 2007; Wankhade et al. 2022) in practical applications.

Consequently, this analysis delves into instances where discrepancies occur, aiming to identify their underlying causes and offer recommendations for improving the precision of automated sentiment analysis. In the context of automatic annotation, the causes of discrepancies appear to be the same for both individuals, indicating that gender does not influence these outcomes.

6.2.2. Causes of Automatic Annotation Issues

When comparing the manual sentiment annotation, where both human annotators reached a consensus, and the sentiment annotation by the flairNLP automated system, a significant disagreement was encountered. The analysis revealed that 163 tweets out of 400 had differing sentiment annotations. The causes for these differences are the following:

- Negations and double negatives: Firstly, automated sentiment detection systems, like flairNLP, can struggle with interpreting negations accurately, leading to misclassification of sentiment. This can be observed in the context of the tweet "@user1 @user2 Trans men have always been men, Individual 1 has never been a woman and is a man" where the tweet was annotated as positive by manual annotators and negative by the automatic system. In this tweet, flairNLP might have focused on the sentence "never been a woman" interpreting the negation as an indication of denying, ultimately annotating it as negative. This misinterpretation can occur because automated systems often rely on negations to comprehend the message without fully understanding the surrounding context or the deeper message conveyed by the text.
  Manual annotators, on the other hand, can recognise that the tweet is reinforcing the identity of trans men and supporting the proper use of pronouns, and the contextual understanding allows them to recognise the intended sentiment as positive, despite the presence of negations.
- Confusion between Subject and Object: Another notable challenge observed in this study is the system's inability to distinguish between the subject and the addressee of the tweets analysed. In the tweet "Given your insistence on being a horrible person, it's clear that understanding the basic concept that he's a man is challenging for you. If that's the case, then it's best not to discuss Individual 1 at all", the system might have interpreted "horrible person" as directed towards Individual 1, leading to a negative annotation. Although the sentiment detection system can classify the overall sentiment correctly based on the semantics of the words, it is not able to discern the direction of the comment or the intended target of criticism. Without the broader context, the system might take the phrase as literal, assuming that it is condemning Individual 1, rather than understanding that it is addressing someone who is misrepresenting or disrespecting him. This underscores the necessity for advanced linguistic models that

can comprehend the context and recognise the relationships between different entities in discourse.
Manual annotators, by contrast, can evaluate the context and correctly interpret that the term "horrible person" is directed toward someone disrespecting Individual 1. This understanding allowed them to see that the tweet's sentiment is, in fact, positive, as it defends Individual 1's identity and advocates for respect.

- Difficulty recognising Sarcasm and Irony: Additionally, automated systems frequently struggle with detecting irony and sarcasm, often leading to misinterpretations in sentiment analysis. For example, in the tweet "Individual 1 transitioned after enduring years of trauma from sexual abuse in Hollywood during her teenage years, following a psychotic breakdown in which she self-harmed, and after experiencing an inner voice urging her to transition [...]", the automated system read this statement as a literal explanation for someone's transition. However, human annotators detected the sarcasm inherent in this comment, understanding that it is questioning or mocking the notion of an "inner voice" leading someone to become trans. As a result, manual annotators classified this as a negative due to the sarcastic undertones, and flairNLP labelled it positive.
- Keyword-based analysis: The last and most prominent cause for discrepancy when employing automated sentiment analysis systems to annotate a corpus is the reliance on keyword-based analysis to classify the sentiment. This approach examines specific words and phrases to determine whether the sentiment is positive, negative or neutral. While this method can be effective for simple cases, it often fails to capture the broader context, emotional subtleties or implicit meanings that human annotators can discern. As a result, discrepancies between human annotators and these systems may arise.
For example, the tweet "@user3 (Individual 1's deadname) was a talented, inspiring and beautiful young woman. Individual 1 is now a disturbing, depressed ghost of their former self." was labelled as positive by flairNLP and negative by the manual annotators. The cause might have been that the automatic system employed keyword-based analysis, identifying words like "beautiful", "talented" and "inspiring" as indicators of positive sentiment. As a result of this focus, the system denied the derogatory use of terms such as "depressed ghost", which eventually led to a positive annotation rather than a negative.

To summarise, automated sentiment analysis systems face significant challenges in accurately detecting sentiment, particularly in online discourse where misgendering is used against TGNC individuals. The present analysis shows discrepancies between automated and manual annotations, which reveal limitations in interpreting negations, discerning subject-object, recognising sarcasm and irony and the reliance on keyword-based analysis. Addressing these issues is imperative to improve the accuracy and reliability of automated sentiment analysis systems such as flairNLP, for future advancements in research.

6.2.3. Wordlist Frequencies

Lastly, to further explore what other forms of discriminatory language co-occur together with misgendering, comparative frequency analyses were performed using the Sketch Engine tool. It becomes possible to identify the positive or negative terms that co-occur with misgendering and their connotations by compiling a list of all unique words in the corpus sample and analysing their occurrences together with their sentiment polarity annotation. This insight is crucial for identifying which aspects related to TGNC identities receive more or less emphasis in the discourse.

To start analysing the sub-corpus targeting Individual 1, the most frequently used unique words extracted using Sketch Engine were "lesbian" (11), "tit" (9), "topless" (5), "mastectomy" (6) and "tittie" (5). To determine whether these terms were being used in a derogatory and harmful context when referring to a trans man, the following table illustrates the frequency of each term's usage and the sentiment assigned to the tweet in which it appears (see Table 8).

Table 8. Derogatory terms targeting Individual 1.

Female-Related Terms	Sentiment Polarity			
	Positive	**Negative**	**Neutral**	**Total**
Lesbian	2	7	1	10
Tit	1	8	0	9
Mastectomy	0	6	0	6
Topless	1	3	0	4
Tittie	2	3	0	5
Total	6	27	1	34

Firstly, the word "lesbian" appears 11 times targeting Individual 1. Most of these instances have negative connotations, suggesting a pejorative or derogatory tone, probably to denigrate the individual's identity or to create confusion between sexual orientation and gender identity. Other terms such as "tits" and "titties" are also frequently used in tweets annotated as negative, suggesting a tendency to objectify or feminise the individual. As for the term "mastectomy", all tweets that include this word show negative sentiment, which may indicate a usage that focuses on gender transition surgeries as an inappropriate process. This reinforces the misperception of Individual 1's gender.

Continuing, when analysing the list of keywords in the sub-corpus targeting Individual 2, the most frequently used unique words were "pretend" (10), "gay" (8), "dude" (6) and "manly" (3). To establish whether these terms were being used in a harmful context when referring to a trans woman, the following table illustrates the frequency of each term's usage and the sentiment assigned to the tweet in which it appears (see Table 9).

Table 9. Derogatory terms targeting Individual 2.

Male-Related Terms	Sentiment Polarity			
	Positive	**Negative**	**Neutral**	**Total**
Pretend	2	8	0	10
Gay	0	7	1	8
Dude	0	6	0	6
Manly	0	3	0	3
Total	2	24	1	27

The term "pretend" is employed 10 times, 8 with a negative sentiment. This suggests an attempt to invalidate Individual 2's gender identity by implying that she is not a woman but merely pretending to be one. The use of the term "gay", when directed at a trans woman, may confuse the interpretation of her gender identity, equating it with sexual orientation rather than recognising her as a woman. Other terms, such as "dude" and "manly", are explicitly masculine terms and present exclusively negative sentiment. These terms are likely to confuse Individual 2 by attributing masculine traits or identities to her, thus denying her correct gender identity.

Overall, the analysis reveals a pattern of discriminatory language strongly connected with intentional misgendering, reflecting deep biases and harmful stereotypes present within online discourse targeting TGNC individuals. Understanding these patterns is crucial for driving systemic change. This calls for a revaluation of how TGNC identities are represented on social media and the wider societal attitudes that support misgendering and discriminatory language. In addition, platforms must take responsibility for fostering inclusive and respectful discourse and implementing mechanisms to identify and address harmful language.

7. Conclusions and Future Research

In this section, a comprehensive examination of the conclusions is derived from the research questions initially posed in the introduction of this study. These research questions are addressed, focusing on understanding the linguistic phenomenon of online intentional misgendering.

7.1. Research Questions

The present study has focused on the analysis of the linguistic phenomenon of misgendering by compiling a corpus sample composed of 400 tweets addressed to two TGNC individuals, with an equal distribution of 200 tweets referring to Individual 1 and 200 referring to Individual 2 and a total of 14,492 tokens and 12,284 unique types. Subsequently, this corpus was manually annotated, and the consensual annotation between two annotators was compared with an automatic system to establish certain issues that may hinder the detection of the phenomenon. After this analysis, the questions posed at the beginning of this study are answered.

- RQ1: Does intentional misgendering as a form of microaggression perpetuate discrimination towards the TGNC community?
  The findings of this study confirm that intentional misgendering significantly perpetuates discrimination against the TGNC community. The analysis and data extracted indicate a significant correlation between intentional misgendering and negative sentiment, suggesting that misgendering indeed contributes to discrimination towards TGNC individuals. The prevalence of negative sentiment associated with misgendering underscores its role in perpetuating discrimination within online discourse.
- RQ2: Does intentional misgendering typically co-occur with other forms of aggression or discriminatory language?
  The study substantiates that intentional misgendering frequently co-occurs with other forms of aggressive or discriminatory language. The analysis reveals that intentional misgendering often accompanies other forms of discriminatory language, such as derogatory terms or negative stereotypes. The co-occurrence of misgendering with such language suggests a broader pattern of discrimination and hostility online towards TGNC individuals.
- RQ3: Is there a significant relationship between the presence of misgendering in tweets and their sentiment polarity?
  The present study's findings strongly indicate a significant relationship between the presence of misgendering in tweets and their negative polarity. Both mispronouning (using incorrect pronouns) and mislabelling (using incorrect gender terms) consistently show a bias towards negative sentiment. This pattern of negativity is evident for both Individuals 1 and 2, indicating a consistent correlation between misgendering and negative sentiment across different contexts.
  Specifically, for Individual 1, tweets containing mispronouning predominantly exhibit negative sentiment, with a significant majority (70 tweets) annotated as negative out of 153 total negative messages. The same applies to tweets with mislabelling (33 tweets), further highlighting the correlation between misgendering and negative sentiment in this context. Similarly, for Individual 2, tweets, including mispronouning and mislabelling, lean towards negative sentiment. Thus, 78 out of 79 tweets with mispronouning were annotated as negative, and 27 out of 28 tweets with mislabelling were annotated as negatives, emphasising a strong association between misgendering and negative polarity.
  Overall, the study's data support the conclusion that misgendering in tweets is significantly associated with negative sentiment, with 208 tweets with misgendering out of the 279 total annotated as negative. This underscores the importance of further exploration into the underlying reasons behind this correlation and its implications for TGNC individuals online.

- RQ4: Can automatic sentiment detection systems effectively identify tweets containing misgendering and expressing hatred towards transgender individuals, or is there a gap in their ability to detect this type of message?
  Automatic sentiment detection systems, such as flairNLP, face inherent limitations that result in the miscategorisation of tweets concerning their overall positivity or negativity. While these systems can sometimes correctly identify positive or negative sentiment, their broader issue lies in a contextual misunderstanding and a lack of nuance in sentiment analysis. This miscategorisation affects the system's ability to accurately flag harmful language, including misgendering, as it struggles to correctly interpret the context in which certain words or phrases are used.
  One of the main limitations is that they rely on keyword analysis to measure the message's sentiment. This approach often ignores the context in which positive or negative terms are used. For example, the presence of positive words in a tweet does not necessarily indicate an overall positive sentiment, as these terms may be used to refer to different persons which complicates accurate detection. In addition, these systems can not identify the addressee or subject of the sentiments leading to erroneous annotations. They operate without context concerning the individuals or groups mentioned and fail to recognise instances where seemingly harmless messages may harm others.
  Furthermore, these automated systems are inadequate at capturing the subtleties of language, including forms such as sarcasm and irony, as evidenced in this study. These linguistic nuances are often crucial in determining the true sentiment and intent of a message, but automated systems have difficulty interpreting them accurately. Hence, without the ability to grasp these subtleties, automatic systems may misclassify the sentiment of a message, leading to inaccuracies in their analysis.
  In summary, automatic sentiment detection systems face significant complications in effectively identifying tweets containing misgendering and expressions of hatred towards transgender individuals. Their reliance on keyword analysis, with a limited understanding of contextual nuances and linguistic subtleties, underscores the need for further development and refinement to enhance their accuracy in detecting and addressing this form of harmful language in online discourse.

To conclude, the present study demonstrates the prevalence of misgendering towards transgender and gender non-conforming (TGNC) individuals, particularly in the context of interactions on the social media platform X. The results reveal that intentional misgendering perpetuates discrimination towards the TGNC community and is not employed intermittently; rather, it is dominant and is accompanied by other derogatory terms that perpetuate discrimination and hostility towards this community. Hence, social media platforms must implement stricter policies and protections for TGNC individuals to foster a more inclusive online environment.

7.2. Future Lines of Research

This study calls for the implementation of robust policies by social media platforms to protect TGNC users, the development of more sophisticated natural language processing tools to better detect and address misgendering, and continued research of the linguistic and social factors contributing to this form of discrimination. Therefore, addressing these areas can create a safer and more inclusive digital environment for TGNC people, promoting their well-being and affirming their online and offline identities.

Additionally, future research could benefit from employing larger datasets to replicate and expand upon these findings. Larger samples would enable more robust analyses and enhance the generalisability of the results, providing deeper insights into the nuances of misgendering and other forms of discrimination across varied contexts.

Other future lines include the development of improved automatic sentiment detection systems which will be used for the identification of misgendering and other subtle forms of discrimination. This entails refining the corpus sample created for this study to improve

contextual understanding and the ability to detect linguistic subtleties. Additionally, future research should explore how linguistic theories such as Context Theory, Inferential Pragmatics, Interactional Pragmatics, and Irony Theory can enhance the development of more sophisticated automatic detection tools. These theories can help create systems capable of understanding the complexities of language, such as sarcasm and irony, thus improving the detection and mitigation of misgendering and other forms of subtle discrimination in online interactions.

Moreover, future research should integrate insights from variational linguistics and sociolinguistics to further refine detection systems. Examining how language varies across different social groups, regions and contexts, offers valuable perspectives on misgendering and other subtle forms of discrimination. By incorporating these insights, automatic detection systems can be adapted to recognise and address diverse linguistic expressions of discrimination more effectively. This approach will enhance the ability of detection tools to operate in varied sociolinguistic contexts, leading to more accurate and contextually aware systems. Ultimately, this could contribute to creating safer and more inclusive digital environments for TGNC individuals, acknowledging and addressing the complexities of language-based discrimination in a more comprehensive manner.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The research data for this study consists of publicly available posts on the X platform. However, the research corpus itself is kept private following X's privacy policies. Access to these messages is governed by the terms and conditions established by X.

Acknowledgments: I would like to thank Borja Navarro Colorado and Victoria Guillen-Nieto for their invaluable support and guidance during my Master's thesis. This article extends the work from my thesis, "She'll never be a man: A corpus-based analysis of misgendering discrimination," which was completed under their supervision in the Master's program in English and Spanish for Specific Purposes at the University of Alicante.

Conflicts of Interest: The authors declare no conflict of interest.

Notes

1 https://x.com/ (accessed on 1 April 2024).

2 Throughout this study, and to ensure the anonymity of individuals targeted by the tweets, any reference to a user on X will be replaced with the placeholder @user and a number.

3 https://www.sketchengine.eu/ (accessed on 25 April 2024).

4 The English Web Corpus (enTenTen) is an English corpus of texts collected from the Internet. The most recent version of the enTenTen21 corpus consists of 52 billion words.

5 SemEval is a series of international natural language processing (NLP) research workshops aiming to further develop the state of the art in semantic analysis by assisting in creating high-quality annotated datasets on an increasingly difficult set of natural language semantics problems. For further information: https://semeval.github.io/ (accessed on 14 April 2024).

References

Akbik, Alan, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. Paper presented at the NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), Minneapolis, MN, USA, June 2–7, pp. 54–59. [CrossRef]

American Psychological Association. 2015. Guidelines for psychological practice with transgender and gender nonconforming people. *American Psychologist* 70: 832–64. [CrossRef]

Argyriou, Konstantinos. 2021. Misgendering as epistemic injustice: A queer sts approach. *Las Torres de Lucca: Revista Internacional de Filosofía Política* 10: 71–82. [CrossRef]

Artstein, Ron, and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics* 34: 555–96. [CrossRef]

Assimakopoulos, Stavros, Rachel Vella Muskat, Lonneke van der Plas, and Albert Gatt. 2020. Annotating for hate speech: The maneco corpus and some input from critical discourse analysis. Paper presented at the Twelfth Language Resources and Evaluation Conference, Marseille, France, May 11–16. Paris: European Language Resources Association, pp. 5088–97.

Biber, Douglas. 1993. Representativeness in corpus design. *Literary and Linguistic Computing* 8: 243–57. [CrossRef]

Birjali, Marouane, Mohammed Kasri, and Abderrahim Beni-Hssane. 2021. A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems* 226: 107134. [CrossRef]

Burghardt, Manuel. 2015. Introduction to tools and methods for the analysis of twitter data. *10plus1: Living Linguistics* 1: 74–91.

Chang, Tiffany K., and Y. Barry Chung. 2015. Transgender microaggressions: Complexity of the heterogeneity of transgender identities. *Journal of LGBT Issues in Counseling* 9: 217–34. [CrossRef]

Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20: 37–46. [CrossRef]

Edmonds, David, and Marco Pino. 2023. Designedly intentional misgendering in social interaction: A conversation analytic account. *Feminism and Psychology* 33: 668–91. [CrossRef]

Fassinger, Ruth E., and Jean R. Arseneau. 2007. "i'd rather get wet than be under that umbrella": Differentiating the experiences and identities of lesbian, gay, bisexual, and transgender people. In *Handbook of Counseling and Psychotherapy with Lesbian, Gay, Bisexual, and Transgender Clients*, 2nd ed. Edited by Kathleen J. Bieschke, Ruperto M. Perez and Kurt A. DeBord. Washington, DC: American Psychological Association, pp. 19–49. [CrossRef]

Guillén-Nieto, Victoria. 2022. Language as evidence in workplace harassment. *Corela* HS-36: 1–21. [CrossRef]

Guillén-Nieto, Victoria. 2023. *Hate Speech: Linguistic Perspectives*. Berlin and Boston: De Gruyter Mouton. [CrossRef]

Havens, Laura, Melissa Terras, Benjamin Bach, and Belinda Alex. 2022. Uncertainty and inclusivity in gender bias annotation: An annotation taxonomy and annotated datasets of british english text. Paper presented at the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP), Seattle, WA, USA, July 15. Stroudsburg: Association for Computational Linguistics, pp. 30–57. [CrossRef]

Hochdorn, Alexander, Vicente Paulo Faleiros, Bruno Camargo, and Paul F. Cottone. 2016. Talking gender: How (con)text shapes gender—The discursive positioning of transgender people in prison, work and private settings. *International Journal of Transgenderism* 17: 212–29. [CrossRef]

Kozareva, Zornitsa, Borja Navarro, Salvador Vázquez, and Andrés Montoyo. 2007. UA-ZBSA: A headline emotion classification through web information. Paper presented at the Fourth International Workshop on Semantic Evaluations (SemEval-2007), Prague, Czech Republic, June 23–24. Stroudsburg: Association for Computational Linguistics, pp. 334–37.

Leymann, Heinz. 1990. Mobbing and psychological terror at workplace. *Violence and Victims* 5: 119–26. [CrossRef] [PubMed]

McCarthy, Linda. 2003. What about the "t"? is multicultural education ready to address transgender issues? *Multicultural Perspectives* 5: 46–48. [CrossRef]

McLemore, Kevin A. 2016. A minority stress perspective on transgender individuals' experiences with misgendering. *Stigma and Health* 2: 1–46. [CrossRef]

McNamarah, Chan Tov. 2021. Misgendering. *California Law Review* 109: 2227–322. [CrossRef]

Moreno-Ortiz, Antonio, and Miguel García-Gámez. 2022. Corpus annotation and analysis of sarcasm on twitter: #catsmovie vs. #theriseofskywalker. *ATLANTIS: Journal of the Spanish Association of Anglo-American Studies* 44: 186–207. [CrossRef]

Nadal, Kevin L., Anneliese Skolnik, and Yinglee Wong. 2012. Interpersonal and systemic microaggressions toward transgender people: Implications for counseling. *Journal of LGBTQ Issues in Counseling* 6: 55–82. [CrossRef]

Nadal, Kevin L., Casey N. Whitman, Lindsey S. Davis, Tania Erazo, and Katherine C. Davidoff. 2016. Microaggressions toward lesbian, gay, bisexual, transgender, queer, and genderqueer people: A review of the literature. *The Journal of Sex Research* 53: 488–508. [CrossRef] [PubMed]

Nadal, Kevin L., David P. Rivera, and Melissa J. Corpus. 2010. Sexual orientation and transgender microaggressions in everyday life: Experiences of lesbians, gays, bisexuals, and transgender individuals. In *Microaggressions and Marginality: Manifestation, Dynamics, and Impact*. Edited by Derald Wing Sue. New York: Wiley, pp. 217–40.

Paludi, Michele A. 2012. *Managing Diversity in Today's Workplace: Strategies for Employees and Employers*. Santa Barbara: Preager.

Pierce, Chester M. 1970. Offensive mechanisms. In *The Black Seventies*. Boston: Porter Sargent, pp. 265–82.

Rosenthal, Sara, Preslav Nakov, Svetlana Kiritchenko, Saif Mohammad, Alan Ritter, and Veselin Stoyanov. 2015. SemEval-2015 Task 10: Sentiment Analysis in Twitter. Paper presented at the 9th International Workshop on Semantic Evaluation (SemEval 2015), Denver, CO, USA, June 4–5. Stroudsburg: Association for Computational Linguistics, pp. 451–63. [CrossRef]

Solórzano, Daniel, Miguel Ceja, and Tara Yosso. 2000. Critical race theory, racial microaggressions, and campus racial climate: The experiences of african american college students. *Journal of Negro Education* 69: 60–73.

Suchomel, Vít. 2020. Better Web Corpora for Corpus Linguistics and NLP. Doctoral thesis, Masarykova Univerzita, Brno, Czech Republic.

Sue, Derald Wing. 2010. *Microaggressions in Everyday Life: Race, Gender, and Sexual Orientation*. Hoboken: Wiley.

Sue, Derald Wing, and Christina M. Capodilupo. 2008. Racial, gender, and sexual orientation microaggressions: Implications for counseling and psychotherapy. In *Counseling the Culturally Diverse: Theory and Practice*, 5th ed. Hoboken: Wiley, pp. 105–30.

Sue, Derald Wing, Christina M. Capodilupo, and Aisha M. B. Holder. 2008. Racial microaggressions in the life experience of black americans. *Professional Psychology: Research and Practice* 39: 329–36. [CrossRef]

Thál, Jakub, and Iris Elmerot. 2022. Unseen gender: Misgendering of transgender individuals in czech. In *The Grammar of Hate: Morphosyntactic Features of Hateful, Aggressive, and Dehumanizing Discourse*. Edited by Natalia Knoblock. Cambridge: Cambridge University Press. pp. 97–117. [CrossRef]

Wankhade, Mayur, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. 2022. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review* 55: 5731–80. [CrossRef]

X. 2024. Abusive behavior. X Help Centre. Available online: https://help.x.com/en/rules-and-policies/abusive-behavior (accessed on 1 April 2024).

MDPI AG
Grosspeteranlage 5
4052 Basel
Switzerland
Tel.: +41 61 683 77 34

Languages Editorial Office
E-mail: languages@mdpi.com
www.mdpi.com/journal/languages

Academic Open
Access Publishing

mdpi.com

ISBN 978-3-7258-3932-2