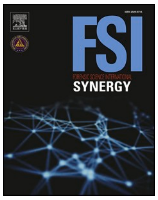
Contents lists available at ScienceDirect

Forensic Science International: Synergy

journal homepage: www.sciencedirect.com/journal/forensic-science-international-synergy



Stylometry and forensic science: A literature review

Valentina Cammarota [a,*], Silvia Bozza [a,b], Claude-Alain Roten [c], Franco Taroni [a]

[a] *School of Criminal Justice, University of Lausanne, Lausanne, Switzerland*
[b] *Department of Economics, Ca' Foscari University of Venice, Venice, Italy*
[c] *OrphAnalytics SA, Vevey, Switzerland*

ARTICLE INFO

Keywords:
Forensic stylistics
Stylometry
Authorship attribution

ABSTRACT

The article focuses on a careful description of literature on stylometry and on its potential use in forensic science. The state of the art of stylometry is summarized to illustrate the history and the scientific foundation of this discipline. However, the study conducted reveals that there are still some key unresolved aspects that require a response from the academic world. The paper introduces the readers to those issues that need to be tackled for stylometry to be accepted as a forensic discipline. In particular, a coherent probabilistic procedure to assess the probative value of the results obtained through this methodology is largely absent. This gap should be filled properly by applying criteria recommended by international organizations such as the European Network of Forensic Science Institutes. Solutions do exist and will allow a better integration of stylometry in forensic science, favouring the acceptance of this scientific technical method in judicial proceedings.

1. Introduction

Over the last decade, there has been a growing interest in the statistical analysis of writing style, otherwise known as "stylometry" [1]. Central to the entire discipline is the idea that everyone has a different writing style, and although it varies over time, stylometry offer extremely valuable information for addressing authorship issues [2,3]. Style would thus represent a so-called and perceived identification means for forensic purposes [4].

Document analysis in forensic science is usually based on the study of a person's handwriting, or on the physico-chemical analysis of the support (generally, paper), the scripting instrument and/or possible printing characteristics [5]. The questions that arise in such a field may concern the origin, production, integrity or legitimacy of the written document [6].

For a long time, the experience of the mandating expert, upon which their opinions were based, represented the main criteria for a scientific report conclusion. Then, quantification tools were favoured for the description and, consequently, the evaluation of a selected list of features con-sidered of relevance for the characterization of the questioned document. This allows not to limit the expert's conclusions mainly to his personal experience [7]. A recent synthesis of the evolution of the state of the art in the field of document and signatures analyses towards quantification is provided by Gaborini [8] and Linden [9,10], respectively.

Preprint submitted to Elsevier May 30, 2024

The ever-growing interest in features description through quantification makes stylometry a promising technique to be adopted in forensic science. Stylometry, a sub-category of the forensic linguistic [11–13], is characterized by descriptive statistical studies of an author's writing style. A text is composed of words (semantic elements), which are organized in sentences (syntax). As the order of the nitrogenous bases determines one's genetic code, the syntax determines the author's style [14]. Textual authorship description takes advantage of discriminating stylistic features [15,16], typically chosen unconsciously by the writer of a text [17].

The contribution of forensic linguistics - a term coined by Svartvik in 1968 [12] - can be useful in forensic investigations [18–21] helping to: (1) identify the author of a text, the language or the speaker; (2) establish the relationship between texts; (3) classify the type of text; and (4) characterize the profile of the author of a text. Stylometry has therefore attracted the interest of the forensic community for a number of reasons, in addition to its ability to meet the need to 'objectify' analyses in the context of documentary forensics [22,23]. Firstly, it can be used to corroborate, or not, the results obtained by the traditional forensic examination, but also to answer questions that the latter was unable to answer. However, the use of stylometry as a forensic method

[*] Corresponding author. École des Sciences Criminelles UNIL – Batochime, CH-1015, Lausanne, Dorigny, Switzerland.
E-mail address: valentina.cammarota@unil.ch (V. Cammarota).

https://doi.org/10.1016/j.fsisyn.2024.100481
Received 22 April 2024; Received in revised form 30 May 2024; Accepted 7 June 2024

in legal cases is currently very limited. Several texts report cases in which the admissibility of stylometric evidence has been questioned and rejected [19,24–26].

Two main steps characterize the stylometric procedure [17]: (1) the selection and the extraction of what are considered as relevant features to represent style markers, and (2) the statistical analysis of the collected data.

Information on the state of the art of stylometry is so scattered that its follow-up can result in a challenging task. Some of a large number of publications summarize the historical development of stylometry techniques [3,23,27,28]. More than a thousand features have been identified as relevant style markers [3] and countless statistical approaches have been implemented. As highlighted in the late '90s [29], there has been no agreement either on which is best markers to select and on the statistical models for data analysis [3,30].

Stylometry is characterized by numerous challenges and developments. This is demonstrated by the existence of PAN, an organization sharing a series of information on scientific events (work-shops and conferences) and digital texts for forensic and stylometry applications. Documents are available at https://pan.webis.de.

This paper will structure information disseminated through a large scientific literature. Section 2 reports the historical development of the domain. Sections 3 and 4 will focus on style markers description and selection and on the most popular statistical approaches for data analysis. Then, sections 5 and 6will present the scientific foundations of this domain and its link with forensic science with an emphasis on challenges for stylometry to guarantee the acceptance of this technical method in judicial proceedings. A conclusion will end this literature review.

2. The origin of stylometry and its evolution

The earliest quantitative descriptions of texts are cited by Rudman [3]. The first, dated back to 300 BCE, was made by Saunaka with reference to an ancient Indian religious text known as the Rig-Veda. Subsequently, Aristarchus of Samothrace, around 180 BCE, recorded expressions which occur either rarely or only once in Greek texts. The writer Aaron Ben-Asher, in 950 AD, counted

different elements, including the number of letters, words, and sentences appearing in the Bible. The word "stylometry" was firstly used by Wincenty Lutoslawski in his work on the chronological tracing of Plato's dialogues [31]. However, the origins of stylometry as a tool for attributing authorship to a text dated back to 1851, with the study of Augustus de Morgan [3,27,28,32], who identified word length as a good style marker. This work motivated Thomas Mendenhall to explore further the potential of this marker [23,28,32–34]. Mendenhall is best known for his 1901 study of the style of plays attributed to Shakespeare, where the aim was to discuss their authenticity [3,35–37]. Many researchers were focused on developing knowledge in stylometry; the most notorious studies have been those of George K. Zipf and George Yule.

Zipf, in 1932, showed a relationship between the number (i) of occurrences of *types* (V) and their frequency (V_i) [23,28]. *Types* (V) are the total number of words in a text that do not repeat, otherwise also called the total number of different words that make up the text [15,23,38]. V_i is so defined as the total number of vocabulary items that are repeated i times. What is known as the "Zipf's First Law" [39] argues that if a list ordered by frequency of occurrence of words in a text is made, a dependency relationship is highlighted between the frequency of words and their position on the list [40–44]. Mathematically, this is described as (1):

$$tfr, i = \frac{\mathrm{c}}{\mathrm{r}_i^{\alpha}} \tag{1}$$

where $tf_{r,i}$ represents the *type* frequency in ith rank; c is a normalization constant for the *Corpus*; and r^{α} the rank (r_i) to the α power, where α was originally ~ 1 and subsequently fitted to empirical data.

Yule, in 1938, identified the instability of sentence length as a marker, highlighting the diffi-culty of defining what a sentence is [23]. Better known is the "Yule (or K) characteristic" devel-oped in 1944;

K characteristic is a metric associated with words occurrence distribution (considered to be Poisson distributed) to describe how often words occur in a text. Many years later Gustav Herdan and Herbert S. Sichel discussed this marker and the adequacy of the Poisson distribution to model probabilistically the words' occurrence in a text [23]. The K-value is calculated using the following formula (2):

$$K = \frac{10^4\ (\Sigma_r r^2 V_r\ -\ N\ )}{\mathrm{N}^2} \tag{2}$$

where r represents the number of repetitions[1]; V_r the number of *types* that appear r times in the text; and N the total number of words in a text (*tokens*) [15,38]. The limits of applicability of the K metric were examined by Baayen [45].

In 1963, thanks to Mosteller and Wallace, an inferential method based on a Bayesian statistical multivariate approach was published [32,46–48]. This evaluative statistical methodology taking advantage of the Mendenhall-based marker choice was implemented to approach (probabilistically) authorship attribution on various American political texts that constitute the Federalist Papers. The choice of this marker together with the probabilistic approach has highlighted the potential offered by stylometry in the practice [23,27,28,49–52]. Among a list of relevant features, functional words [29], including synonyms (a marker introduced by Ellegard two years earlier), were considered of extreme importance for the field. Thirty years later, this work was analysed us-ing a metric called "character *n*-grams" by Ref. [53] (a definition of character *n*-grams will be provided in Section 3), but unfortunately the probabilistic approach suggested by Mosteller and Wallace was apparently unreasonably discharged.

From 1964 until the end of the '90s, multiple studies were carried out in order to search style markers and to develop computer-assisted textual authorship description methodologies [34]. Starting from the '70s, technological development, accompanied by the increasing availability of software facilities and of texts in electronic format, favoured a sensible reduction in processing time [52], but also an increase in the number of published studies [23]. Subsequently, starting from the '90s, a strong impulse to stylometric analysis was growing by the diffusion of the ma-chine learning (ML) techniques for description and classification purposes [54].

3. Style markers

As mentioned above, the first step of any stylometric procedure is based on the selection and extraction of stylistic features. Numerous, these features are categorized in various ways in the literature. The following is a classification of markers that attempts to take into account the various aspects presented in other studies.

3.1. Classification of style markers

Some features are considered as class characteristics, others are described as individual char-acteristics [13]. The first kind of features can be valuable to restrict the circle of potential authors to a specific population. They are related to external factors describing a given group, such as the social class. The second one refers to one's personal development of the language and its use; they are classified as linguistic features whose rarity is such that the probability that they could be reproduced by a third party is considered negligible.

3.1.1. Lexicon

A text can be conceived as a sequence of elements, called *tokens*

[1] Please the reader may be careful: the meaning of r in the two formulae (1) and (2) is not the same.

(words, numbers, punctuation marks) grouped in sentences [34]. Historically, the first works on stylometry were based on these elements, which are defined as style markers at the lexical level. Among the most frequent that can be listed: word length [15,23,55,56], sentence length [15,23], syllables [23,57], discourse elements [23], function and content words [23,27,28,30,58,59], and vocabulary richness[2] [23,28,35,39,60–71]. Savoy [41] mentions a further marker that is less common in the literature: the relative frequency of "long words" (words composed of more than 6 characters). The higher the relative frequency, the more sophisticated and complex the style.

Six main drawbacks of lexical markers have been identified: (a) the ease of fooling the system [30]; (b) the influence of editors, who can intervene and have an impact over the presence or absence of a word [30]; (c) the failure to take into account morphologically related words [30]; (d) the difficulty, or impossibility, of determining similarities concerning alternative word forms, probably due to usage errors [58]; (e) the difficulty in defining what is meant by "word" which is not necessarily easy in some languages [58,72]; (f) the dependence on the typology of the text being written, as well as on the context [23].

It's also worth noting that a different study [73] showed the importance of certain words at the level of text structure and not only with respect to frequencies and syntactic relations; their use is restricted by the length of the texts under investigation, which must be sufficiently long.

3.1.2. Character

The second type of marker refers to the character level. In this case, the text is treated as a sequence of characters. More precisely, each text is represented as a frequency vector of selected character sequences. In other words, the text is thus considered to be a "bag-of-character" [74]. Historically, the study of characters in stylometry is linked to Leon Battista Alberti, whose aim was to distinguish prose from poetry in Latin works [41].

The specific character sequences chosen are called *n*-grams, where *n* represents the number of characters that make up the selected sequences. The use of *n*-grams has been introduced in the context of the characterization of poets [75]. This approach has also been used to highlight the respective contributions of several authors of a given text [46].

As this marker is independent of the language of the text,[3] its use is particularly advantageous [17,28,34,72]. With reference to the limitation of lexical markers (*e*) mentioned above, the use of the *n*-gram is particularly interesting in cases where it is difficult to define the 'word', as in Asian languages (e.g., Chinese, Japanese, Korean). Furthermore, the observation of *n*-grams is also applicable in DNA sequence studies, music field [17,72], as well as for the characterization of digital code sources [76].

The use of *n*-grams is also favoured by the limited need to pre-process the text[4] [80] and by the fact that they can be easily extracted and counted [58]. Furthermore, another advantage is the small impact of noise, possibly caused by language mistakes or misprints [58].

Since altering a style at the character level is more difficult than at the lexical level, the use of this marker is to be preferred in cases where there is a suspicion of active malice [28], notably a deliberate alteration of writing style for malicious purposes.

Leaving aside the ease of extracting this type of marker [28], its implementation in the practice is hampered by two concrete difficulties: the choice of the character sequence length (*n*) and the length of the text profile size (*L*). By definition [80]:

> A profile *P* is a set of *L* pairs (g_1, f_1), (g_2, f_2), …, (g_L, f_L), where g_1, g_2, …, g_L are the *L* most frequent n-grams of the text (in decreasing order) and f_1, f_2, …, f_L their normalized (wrt text length) frequencies of occurrence, respectively (at p.237),

and the optimal character sequence length (*n*) was found to be language dependent, as the average word length can vary [17,28,46]. In the literature there can also find proposals according to which a combination of, say, 2-g, 3-g or even more is to be preferred to a single value for *n* [17,28,46]. Clearly, this leads to an increase in dimensionality requiring features reduction methods. Another approach would be to take into account only the dominant *n*-grams as suggested by Houvardas and Stamatatos [46]:

> The main idea is to compare each n-gram with similar n-grams (either longer or shorter) and keep the dominant n-grams. (at p.79)

3.1.3. SYNTACTIC elements

The idea that every author uses, unconsciously, the same syntactic pattern leads to the identifi-cation of this third category of markers [34]. Texts can, for example, be broken down into what is called "Part-of--speech" (POS) and then the different parts can be used to search for features such as punctuation [30]. The latter marker is considered very effective in cases where texts have not been subjected to any edited review [77].

Syntactic information as a marker has two important limitations: it is language-dependent and highly vulnerable to analysis errors [28,34]. Despite the fact that syntactic structure can be characterized by high intra-variability [81], it has more reliability than lexical markers [82].

3.1.4. Semantic elements

The study of semantics consists in analysing the context in which words are used. For example, it has been said that in Corneille's comedies, the word "love" is associated with the father figure [41]. As elements that reflect the meaning of texts, semantic elements include subjects [83].

While the markers previously described can be automatically investigated, the complexity of semantic analysis makes its application for stylometric analyses difficult to implement [34].

3.1.5. Specific features

Application-specific features of text, like its structure and layout, anomalies and metadata (i.e. information describing electronic documents) [30], can also be used as stylometric markers. They define the structure of a text and are particularly useful when analysing email texts and web forum discussions [34].

Neal et al. [28] also distinguished so-called "content-specific" features of the text, which in-clude keywords and other information about the topic of the document.

3.1.6. Other textual elements

Faults or words that reflect a social or cultural reality has been suggested as additional markers [28]. Note the difficulty in their classification into one of the above-mentioned typologies of markers.

3.2. Some remarks

As pointed out by Athira and Thampi [84], the choice of the style marker to be targeted is a fundamental step in text analysis. However, to

[2] The vocabulary richness is not necessarily very effective when it comes to problems of authorship, but it is helpful in cases where one wishes to distinguish the writings of a machine from those produced by a human being [41].

[3] In order to avoid misunderstandings, it should be made clear that the *n*-gram model is language-independent. In the sense that no language-dependent adaptations are necessary. Regardless of the language, this marker can be used. However, all texts in the same case must be written in the same language.

[4] It is important to note that the published version is subject to the intervention of editors and others [30], which could lead to changes in the original text [77,78]. Compared to the holograph, the first edition remains the clos-est [3]. Pre-processing is often necessary before the analysis of texts to remove elements that may belong to the reviewers/editors [79].

this day, there is no consensus on which marker should be considered as the best. Traditionally, the choice of stylometric marker was left to the scientist's personal choice [85].

In 2007, a study focused on comparing the effectiveness of 39 style markers was published [15]. The *n*-grams were the markers that achieved the highest accuracy rates. Overall, a de-crease in accuracy could be observed as the number of authors in the *Corpus* (i.e. the set of reference/comparison material) increases. The use of 2-g and 3-g provided the best re-sults. As reported by Grieve [15], previous studies have shown the effectiveness of longer *n*-gram. Grieve points out that this effectiveness may be related to the fact that all analysed texts in previous studies had the same common topic and so longer *n*-grams are therefore more discriminating.

The use of the most frequent words and *n*-grams as markers have provided the most promising results in recent scientific works [86]. As reported by Koppel et al. [51], *n*-grams perform well in the forensic context and have been defined as very sensitive and accurate markers [87]. The choice of selecting and excluding the most frequent words as markers is proposed under the assumption that there are classes of words beyond the author's control, whereas *n*-grams have shown language independence and the possibility of capturing different stylistic information,[5] favouring their use [54].

Stamatatos [58] makes the reader aware that studies are often conducted using extended texts characterized by similarities in the genre and topic. This does not correspond to practical (case-related) conditions observed in forensic science, where the available questioned and reference materials are often extremely limited.

The author points out that controlled conditions in scientific studies inevitably lead to an over-estimation of the effectiveness of the method. By taking texts of different genres and topics, Sta-matatos [58] was able to report the effectiveness of the *n*-grams in comparison to function words (e.g. prepositions, common adverbs, pronouns and articles [59]) and their accuracy.

Style markers can also be used in combination [34,84]. The joint consideration of features leads to an increased amount of information [30], but also in dimensionality, which can have noisy effects affecting parameters estimation procedures due to recurrent characteristics [34,41].

4. Data analysis

By the '90s, multivariate statistical analysis to depict groups of writers was used as an alterna-tive to previous univariate approaches based on the measurement of a single feature [23,30,82,88,89].

It is not the purpose of this paper to exhaustively review and summarize all statistical tech-niques used in the stylometric field. Below, the reader can find a general overview of the most cited methods in the literature [3,23,28,30,35].

A long list of techniques for descriptive and inferential statistical approaches were described for stylometric data analysis since late '60. These techniques refer to Ref. [23]: chi-square test (see, e.g. the works of Brinegar [90]), factor analysis (e.g. Miles and Selvin in 1966), discriminant analysis (e.g. Cox and Brandwood in 1959, Somers in 1966, Bruno in 1974 and Ledger in 1989) and clustering (e.g. Bailey in 1979, Boreland and Galloway in 1979, Ledger in 1989 and Holmes in 1992). The general aim of those proposals is to describe and to detect similarities or dissimilarities between features of relevant markers in different texts.

A criticality that characterized this type of application is linked to the high dimensionality of the available data. Principal component analysis (PCA) was recognized as a good method in the field of stylometry [11] and indeed it has proven to be a viable alternative for dimensionality reduction [30]. Historically, some of the most significant studies relying on this method include the works of Burrows and Hassall, 1988; Burrows, 1992; Smith, 1991 and Holmes, 1992 [23]. PCA was also used to show that the 15th Oz book' writing style is more compatible with that of Thompson's than that of Baum, author of the earlier Oz books [31]. Note that some criticisms on the use of PCA in genetics studies have been published recently [91].[6]

In the '90s, some studies based on correspondence analyses (a multivariate technique used to describe potential relationships through graphical representations [92]) has been published. In the same years, A. Q. Morton introduced into the field of textual authorship studies the so-called cumulative sum control chart (CUSUM), a statistical technique that quantifies the cumulative sums of deviations of the sample values from a target value [32,93]. The validity of this proposal has been seriously questioned by several researchers, mainly because of a lack of clarity both in the interpretation of results and in the understanding assumptions [29,94].

As mentioned at the end of section 2, machine learning (ML) techniques play an important role in stylometry [11,16,28,35,49,53,58,75,95–98]. These techniques provide higher accuracy than descriptive statistical methods, as well as better tolerance to noise and non-linear interactions between different style markers [82]. Some critical remarks on their use have been highlighted by Ref. [99]:

> Once you unleash it on large data, deep learning has its own dynamics, it does its own repair and its own optimization, and it gives you the right results most of the time. But when it doesn't, you don't have a clue about what went wrong and what should be fixed. In particular, you do not know if the fault is in the program, in the method, or because things have changed in the environment. We should be aiming at a different kind of transparency. (at p. 1)

Dimensionality reduction and classification are the main advantage of ML techniques, regard-less of whether they are used in supervised or unsupervised learning environments. However, as shown in earlier studies, the dimensionality of forensic data does not represent a barrier (see, e.g., Refs. [100,101]). As pointed out by Biedermann [102], there are at least two major problems that the implementation of a ML-based approach poses with regard to forensic identification problems, such as the authorship procedure in stylometry. Primary, it is important to recall the purpose of the forensic evaluative process and in particular the role of the expert in the authorship attribution. Secondly, a ML-based approach provides inferential answers based solely on the data used to build the model. What about case-related information and their role in the process? As noticed by, e.g., the Swiss jurisprudence,[7] the expert assess the value of data so to provide a statistical information and it is not superfluous to assess all other evidence in order to answer a question of legal interest and deliver a final conclusion on the hypothesis of interest (i.e., authorship). Moreover, decision theory plays an important role in legal procedures; in fact, a decision maker should coherently justify their conclusion and to do so he should adopt a decision theoretical point of view where in addition to the uncertainty that characterizes the elements of decision-making, there is also the need to specify a preference ordering among the consequences associated with each decision and quantify their undesirability [103,104]. Biedermann [105] criticized the use of ML in forensic science:

> [t]he persistence of source attribution claims [individualization or authorship] in foren- sic science literature is problematic. […] Machine learning not only bears potential for misapplication for the

[5] Characters markers provide lexical, syntactic and/or structural information [46,80].

[6] It has been stated that "Principal Component Analyses (PCA)-based findings in population genetic studies are highly biased and must be reevaluated.", and that "We conclude that PCA may have a biasing role in genetic investi-gations and that 32,000–216,000 genetic studies should be reevaluated." [91].

[7] See, e.g. Bulletin de Jurisprudence Pénale, 4 (1997) 103–105.

> purpose of forensic source attribution, but that such misap- plications actually do occur. (at p.1)

In many scientific or industrial sectors, ML technique can be thought of as an engineering tech-nique that provides valuable tools for optimization, but its use in forensic science is still questioned.

The authorship process requires transparency and robustness in the model and the rational use of contextual case information. The term "transparency" (in the words of Jackson [106])

> [explains] in a clear and explicit way what we have done, why we have done it and how we have arrived at our conclusion. We need to expose the reasoning, the rationale, behind our work. (at p. 84)

The term "robustness" challenges a scientist's ability to explain the ground for their opinion to-gether with their degree of understanding of the particular evidence type. Do the current inferential statistical methods used in practice satisfy those requirements?

Other descriptive methods based on distance or similarity [85] between given texts are also widely implemented to deal with textual authorship attribution [28,30,107,108] but with a ques-tionable line of reasoning: the smaller the distance value, the higher the probability that the texts were written by a given person (author) [38,78,108].

It is worth noting that a forensic scientist should evaluate the available data (e.g. data summarized by similarity scores or distances between features extracted from questioned and reference texts). Once the distance between the questioned texts has been calculated (or a list of features described), the expert should (as requested by the ENFSI guideline for evaluative reporting [109]) assign (at least) two conditional probabilities: the probability of obtaining such measurement un-der the hypothesis that the texts were authored by the same source (author) from one side, and under the hypothesis that the texts were written by different authors, from the other side.

> Evaluation will follow the principles outlined in Guidance note 1 (refer to paragraph 4.0). It is based on the assignment of a likelihood ratio. Reporting practice should con- form to these logical principles. This framework for evaluative reporting applies to all forensic science disciplines. The likelihood ratio measures the strength of support the findings provide to discriminate between propositions of interest. It is scientifically accepted, providing a logically defensible way to deal with inferential reasoning. Other methods (e.g., chemometrical methods) have a place in forensic science, to help answer other questions at different points of the forensic process (e.g., validation of analytical methods, classification/discrimination of substances for investigative or technical reporting). Equally, other methods (e.g., Student's t-test) may contribute to evaluative reports, but they should be used only to characterize the findings and not to assess their strength. Forensic findings as such need to be distinguished from their evaluation in the context of the case. For the latter evaluative part only a likelihood ratio based approach is considered. [109], p. 6

The ratio between these two probabilities is known as the likelihood ratio, or more generally the Bayes factor, the coherent measure for the value of the evidence [110]. In stylometry, this probabilistic approach is rarely implemented or discussed. A partial probabilistic model, referring just to one of the two hypotheses, has been proposed by Ref. [79]. A Bayesian probabilistic approach has been recently proposed in stylometry by Ref. [111] in the context of authorship discrimination between French authors of classical plays.

5. Forensic science and stylometry

Following the themes presented in the previous sections, it is not surprising that forensic science can show an interest in stylometry. Before delving more deeply into the link between the two disciplines, it is necessary to elaborate on the concept of 'textual authorship attribution'. This notion is extended to four principal different situations: identification, verification, profiling and similarity detection.

A problem of authorship attribution may be based on a question of identification or verification. In the first case, the unknown identity of the author is sought. In the second case, the textual origin of different documents is analysed in order to determine whether or not the texts come from the same source [28,71,112]. Generally, in the case of verification the questioned text is compared only with reference texts sourced by a given candidate [41,51]. Identification and verification questions can also be treated as classification cases [113], which in turn can be described as: (i) a binary classification where all the questioned documents will be considered to have been written by one of the two known authors: (ii) a multi-class classification where all the questioned documents come from more than two known authors and; (iii) finally, a one-class classification where there is only one known author and the aim is to determine which document(s) was (were) written by that author [2,12,34,114].[8]

Profiling aims to help determine the personal, demographic or psychological characteristics of the author(s) of a text. Typical questions encountered may concern: gender, age, social class, education level, and nationality of the author(s), as well as the number of individuals behind the writing of a text [30,51,71,108,115,116].

Vice-versa, in the so-called "similarity detection" situations, the scientist's interest is to investigate whether texts were written by a single author without ascertaining their identity and/or inferring their personal characteristics. This is typically what is sought in suspected plagiarism cases [82]. While the software currently implemented in practice has the ability to detect direct plagiarism (copy/paste of each word), the problem is more difficult in the case of so-called "the-saurus plagiarism" characterized by the use of synonyms and a mixture of sentences from different texts. Moreover, difficulties arise when a third party wrote a text by officially substituting their mandating party [117].

Other situations were described by Savoy [41] who further distinguishes between 'author clus-tering' and 'author linking' problems. In the first case, the interest is in detecting the number of authors hidden behind a sample of *n* texts, whereas in the second case one is interested in deter-mining the texts that were written by the same person, while not having any information on the number of writers.

However, it should be emphasized that stylometry is not focused exclusively on problems of authorship attribution [118]. Another branch of stylometry, known by the term "stylochronometry", focuses on problems of a chronological nature (e.g. dating a text) [28,29,35].

Stylometry can play a valuable contribution in providing answers of great interest for the foren-sic field [4], say, did the suspect actually write the text in question? Who, among a group of people, could have written the text in question? Were these documents all written by the same person? When was the text written? What is the gender (or, e.g. the mental state) of the person who wrote the text?

These questions may be encountered on various scenarios; in fact, the applicability of sty-lometric methods is not limited to issues of pseudonymity, plagiarism - professional, literary (especially with the creation of e-books) or musical - or contractual cheating. Other areas of interest that may be subjected to stylometric analysis are included in the non-exhaustive list below: employment contracts, threatening or harassing letters, wills, testimonies, extortion at-tempts, threats of attack (e.g. related to terrorism), suicide letters, malware, cyber harassment, sms messages, emails, blogs (e.g. in the field of child pornography) or, again, defamatory posts [1,20,21,24,28,34,46,74,79,82,84,87,114,117,119].

Stylometry applied to the above-mentioned scenarios may lead to discussions at trials around its admissibility as reliable technique.

[8] It is legitimate to estimate that this categorization is not exhaustive, in fact, it lacks, for instance, the case where several authors may be the source of a text.

5.1. The theoretical foundations of stylometry

If we consider the definition of stylometry as an approach to infer the characteristics or the identity of the author of a text based on discriminating stylistic features [16], the importance of so-called features, otherwise referred to as stylometric markers, is evident. Their main property is that these are typically chosen unconsciously by the writer [17]. More than a few properties were mentioned in Ref. [29]. It is reported that:

> [T]hey should be salient, structural, frequent and easily quantifiable, and relatively immune from conscious control. By measuring and counting these features, stylometrists hope to uncover the 'characteristics' of an author. (at p. 111)

Textual attribution follows from the basic assumptions of stylometry, notably that (a) each individ-ual has a single, verifiable style [3] (therefore two individuals can be discriminated by their style) and that (b) one's style evolves over time [2], as a result of the writer's personal development [98], but it is still discriminating.

Since 1964, styles were defined as deviations from a norm [120]. Unconsciously, each indi-vidual uses a unique way of writing text, which is described by quantifiable markers that allow for a style characterization [27,98,121]. Here is an excerpt from McMenamin's statement in the *Ceglia vs Zuckerberg* expert opinion [122], which sums up the concept:

> Individual differences in writing style are also very often due to an individual's choice of available alternatives within a large, shared common pool of linguistic forms. At any given moment, a writer picks and chooses just those elements of language that will best communicate what he/she wants to say. The writer's 'choice' of available alternate forms is often determined by external conditions and then becomes the unconscious result of habitually using one form instead of another. Individuality in writing style results from a given writer's own unique set of habitual linguistic choices. (at p. 3)

As a result of the above, the assumption concerning the evolution of a person's style over time seems to be plausible. In addition to the effect that certain diseases, such as dementia, may have on writing style, a decrease in the diversity of word usage (also referred to as richness of style) has been observed when considering a writer's style over time [23,123,124]. In this regard [41], explains that 5 life phases can be associated with the language and style of each individual: the infant phase, the child phase, the adolescent phase, the adult phase and the elderly phase. In the adult phase the style undergoes little variation and is relatively stable.

The study by Ref. [124] was able to show the possibility of classifying a text correctly within an interval of time. The authors pointed out, however, that longer periods of time between the writing of two texts do not necessarily imply glaring changes in style.

Although style is unique to each individual, different styles can be adopted by the same person (according to background, text genre, writing period, theme, text form, or audience [113,125]). This implies that there is some variability in personal style, in addition to the variability that might be due to the passage of time. These two aspects recall the two fundamental "laws" of handwriting [126]: the first referring to the inter-variability (no two individuals write exactly alike) and the second relating to the intra-variability (no one person writes the same word the same way twice) of handwriting.

Analogous to cases typically encountered in forensic science, the problem of textual author-ship is addressed by searching for a "stylistic fingerprint" of the questioned text in order to make a comparison with the stylistic fingerprint characterizing the reference material [78]. The set of comparison material is referred to as the *corpus*. Basically, the questioned text is thus compared with texts originating from a given population, which will be specified on the basis of the hypoth-esis set in question [3]. Closed-class scenario refers to the situation where it is assumed that the true author is one of those whose texts make up the *corpus* [71,79,88]. In contrast, the open-class problem admits the possibility that a third party, not belonging to the *corpus*, may be the author of the questioned work [30,127] such as in the "Ferrante case" studied by Mikros [54].

Some publications [3,79] attempt to list the conditions necessary to ensure quality comparison material. Its textual authorship must certainly be known and verified. Following the assumption that time influences the style of each individual, reference material should be as contemporary as possible with the questioned text. Other assumptions, including the dependency of style or literary genres/types of text need to be tested further, though it seems that this may have a negligible impact when certain stylometric markers are used [98].

What must also be tested is the minimum length that the questioned text should have in order to guarantee robust results, as well as the minimum amount of reference material. This question is often mentioned in the literature and represents a real practical problem. Nowadays, with the increasing use of portable media and devices, stylometry is increasingly striving to be applied to shorter and shorter texts. Obviously, in contrast to research studies, real forensic cases are characterized by short texts and a very limited amount of material [16,34,47,58,116,128–130].

5.2. Admissibility

The admissibility of so-called *language evidence* in court was discussed in the early 2000s, notably in *United States vs Van Wyk* [128]. At the heart of the still limited use of stylometry as a scientific methodology to effectively address forensic cases is the lack of scientific studies on the effectiveness and robustness of the methods employed [79,131]. Chaski [128] explains:

> The Defence argued that the 'proffered expert testimony is subjective, unreliable and lacks measurable standards' (Van Wyk, 83 F. Supp.2d 515, 521). Thus, the Defence argued that admitting forensic stylistics testimony would violate several other criteria of the Daubert standard of empirical reliability, such as falsifiability of the technique, known error rate, and standard operating procedures for performing the technique (see Note 2 for more discussion of admissibility factors). (at p. 2)

In fact, one should guarantee that the technique has been (or could be) tested, that the technique has been published in peer-reviews, that the technique is accepted by the scientific community and that its error rate is known [132–134]. A further fundamental aspect relates to the adequacy of the evaluative procedure and the adherence to the published international standards (see, e.g. recom-mendations offered by the ENFSI guidelines [109]) to deal with the assessment of the available descriptive data of a series of texts.

The aspects that mostly limit the use and admissibility of stylometry for forensic purposes are method credibility and potential active malice [30]. The term "credibility" hides different requirements [30]; in order to be credible a method must be accurate and the degree of accuracy must be determined, its scientific foundations must be recognized, and finally the full methodology must be transparent and robust. While this is partially reported in the literature, little is known about the deliberate alteration of writing style for malicious purposes (adversarial stylometry). Three alteration techniques are known: imitation, translation and obfuscation [28]. Savoy [41] points out that imitation is a very difficult task both to accomplish and to detect. The techniques of translation and obfuscation are less complex. The first practice consists of translating the original text into one or more languages and then returning to the original language, while the second one is characterized by the substitution of words following the identification of the stylistic profile that one wishes to change.

Thus, although stylometry has existed for centuries and its use is scientifically documented, its implementation as a scientific method to effectively deal with court cases is still discussed and rather limited.

6. Conclusions

The principal aim of this paper was to conduct a review of the state of the art of stylometry, along with a description of the main historical stages, and most importantly a discussion focused on the link with forensic science, its potential contribution and the eligibility of these tools in court. The persistence of critical issues that must be addressed to allow (or favour) the implementation of these tools in the forensic field is palpable. Open issues include technical and evaluative aspects. There must be included the choice of the marker, the choice of the distance to quantify similarities, or the minimum requirements in terms of quantity and quality of material. However, there are still major gaps to be filled, particularly with regard to the evaluation of the measures collected and for decision-making purposes such as identifying an author, characterizing the author, detecting a false document, detecting plagiarism, the contribution of several authors, etc. Ishihara [135] had already pointed out the failure of forensic stylometry that unlike other forensic fields (including handwriting examination) does not promote (or support) the use of a likelihood ratio for evaluative and investigative issues of forensic interest.

The probabilistic approach proposed by Bozza et al. [111,136] represents a first attempt in this direction. Champod et al. [137], as reported by Riva [138], highlighted the problematic issue of probability assignment. Unfortunately, this aspect is still wrongly considered to be a major obstacle to the wider use of the probabilistic approach.

It must be pointed out that, from the review of the state of the art in the field of stylometry, it emerges quite clearly that the role of the scientist in rarely respected. The scientist is asked to evaluate observations made on the material under examination and on reference materials. This definition is found in many legal texts. The role of the scientist is to evaluate observations and not hypotheses that may be of interest to the Court [139,140]. To fulfil the meaning of terms such as “identification”, “verification” and “profiling” (as usually mentioned in stylometry), it is not sufficient to consider only data collected on available materials. As stated above throughout law jurisprudence, any information gathered during the investigative procedure related to the disputed document is of essential interest to justify a decision about the hypotheses. In other words, the expert must focus exclusively on the value of the observation (i.e. the data obtained through stylometry analysis), and not on the evaluation of the hypotheses (i.e. the text was written by author X). Thinking that an *ad hoc* statistical method (i.e. ML technique, PCA, etc.) solely based on data can solve the legal problem and allow one to express their opinion on an hypothesis is inappropriate. Each actor involved in the criminal proceeding should respect their role. The scientist must be able to quantify the value of the evidence and provide this information to the judge, so that it can be integrated with the mass of information available to the Court in order to reach the final verdict. Note that this should include the quantification of the undesirability of decision outcomes (e.g., a false authorship attribution). Needless to say, the full Bayesian model is the only one that allows one to perform this sequence of evaluation and decisions.

CRediT authorship contribution statement

Valentina Cammarota: Writing – original draft, Conceptualization. **Silvia Bozza:** Writing – review & editing, Validation. **Claude-Alain Roten:** Conceptualization. **Franco Taroni:** Writing – review & editing, Validation, Supervision.

Declaration of competing interest

The authors declare that they have no competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors thank the Swiss National Science Foundation for its support through grant number 100011-204554/1 (The anatomy of forensic inference and decision).

References

[1] A. Abbasi, H. Chen Writeprints, A stylometric approach to identity-level identification and similarity detection in Cyberspace, ACM Trans. Inf. Syst. 26 (7) (2008) 1–7:29.
[2] D. Pavelec, E. Justino, L.S. Oliveira, Author identification using stylometric features, Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial 11 (2007) 59–65.
[3] J. Rudman, Authorship attribution: statistical and Computational methods, in: Keith Brown (Ed.), Encyclo- Pedia of Language & Linguistics, second ed., Elsevier, Oxford, 2006, pp. 611–617.
[4] P. Juola, Future Trends in authorship attribution, in: P. Craiger, S. Shenoi (Eds.), *Advances in Digital Forensics III*, IFIP — the International Federation for Information Processing, Springer, New York, NY, 2007, pp. 119–132.
[5] D. Ellen, S. Day, C. Davies, Scientific Examination of Documents: Methods and Techniques, fourth ed., CRC Press, Boca Raton, 2018.
[6] R.A. Huber, A.M. Headrick, Handwriting Identification: Facts and Fundamentals, CRC Press, Boca Raton, 1999.
[7] R. Marquis, S. Bozza, M. Schmittbuhl, F. Taroni, Handwriting evidence evaluation based on the Shape of characters: application of multivariate likelihood ratios, J. Forensic Sci. 56 (2011) S238–S242.
[8] L. Gaborini, Bayesian models in questioned handwriting and signatures. Université de Lausanne, Faculté de droit, des sciences criminelles et d’administration publique, Lausanne, 2021. PhD thesis.
[9] J. Linden, Forensic examination of dynamic signatures. Université de Lausanne, Faculté de droit, des sciences criminelles et d’administration publique, Lausanne, 2022. PhD thesis.
[10] J. Linden, R. Marquis, S. Bozza, F. Taroni, Dynamic signatures: a review of dynamic feature variation and forensic methodology, Forensic Sci. Int. 291 (2018) 216–229.
[11] A. Clark, Forensic stylometric authorship analysis under the Daubert standard, SSRN 2039824 (2011).
[12] D. Pavelec, L.S. Oliveira, E.J.R. Justino, L.V. Batista, Using Conjunctions and adverbs for author verification, J. UCS 14 (2008) 2967–2981.
[13] G.R. McMenamin, Forensic Linguistics: Advances in Forensic Stylistics, CRC Press, Boca Raton, 2002.
[14] G. Genilloud, C.A. Roten, Determination by Stylometry of the Probable Author of the Ferrante Corpus: Domenico Starnone — OrphAnalytics SA, 2016.
[15] J. Grieve, Quantitative Authorship Attribution: an Evaluation of Techniques, vol. 22, Literary and Linguistic Com- puting, 2007, pp. 251–270.
[16] W. Oliveira, E. Justino, L.S. Oliveira, Comparing Compression models for authorship attribution, Forensic Sci. Int. 228 (2013) 100–104.
[17] V. Keselj, F. Peng, N. Cercone, C. Thomas, N-Gram-Based author profiles for authorship attribution, in: Proceedings of the Conference Pacific Association for Computational Linguistics (PACLING) vol. 3, Dalhousie University Press, Halifax, 2003, pp. 255–264.
[18] J. Olsson, Wordcrime: Solving Crime through Forensic Linguistics, A&C Black, 2009.
[19] C.E. Chaski, Best practices and admissibility of forensic author identification, J. Law Pol. 21 (2) (2013) 333–376. ISSN 1074-0635.
[20] H. Houtman, S. Suryati, The history of forensic linguistics as an assisting tool in the analysis of legal terms, Sriwijaya Law Review 2 (2018) 215–232.
[21] L. Renaut, L. Ascone, J. Longhi, De la trace langagière à l’indice linguistique : Enjeux et précautions d’une linguistique forensique, Ela. Etudes de linguistique appliquee 188 (2017) 423–442.
[22] D. Dessimoz, C. Champod, Linkages between Biometrics and Forensic Science, 2007, pp. 425–459.
[23] D.I. Holmes, Authorship attribution, Comput. Humanit. 28 (1994) 87–106.
[24] C.E. Chaski, Who’s at the Keyboard: authorship attribution in digital evidence investigations, in: Presented at the 8th Biennial Conference on Forensic Linguistics/Language and Law, 2005.
[25] M. Coulthard, On admissible linguistic evidence Symposium, J. Law Pol. 21 (2) (2013) 441–466.
[26] Patrick Juola, The Rowling case: a proposed standard analytic protocol for authorship questions, Digital Scholarship in the Humanities 30 (2015) i100–i113.
[27] D.I. Holmes, J. Kardos, Who was the author? An introduction to stylometry, Chance 16 (2003) 5–8.
[28] T. Neal, K. Sundararajan, A. Fatima, Y. Yan, Y. Xiang, D. Woodard, Surveying stylometry techniques and applications, ACM Comput. Surv. 50 (86) (2017) 1–86: 36.
[29] D.I. Holmes, The evolution of stylometry in Humanities Scholarship, Lit. Ling. Comput. 13 (1998) 111–117.
[30] P. Juola, Authorship Attribution, ume 3, Now Publishers Inc, 2008.
[31] J.N.G. Binongo, Who wrote the 15th book of Oz? An application of multivariate analysis to authorship attribution, Chance 16 (2003) 9–17.
[32] Harold Love, Attributing Authorship: an Introduction, Cambridge University Press, Cambridge, 2002.

[33] T.C. Mendenhall, The characteristic Curves of Composition, Science 9 (1887) 237–249.

[34] E. Stamatatos, A Survey of modern authorship attribution methods, J. Am. Soc. Inf. Sci. Technol. 60 (2009) 538–556.

[35] F. Can, J.M. Patton, Change of writing style with time, Comput. Humanit. 38 (2004) 61–82.

[36] C.B. Williams, Mendenhall's studies of word-length distribution in the works of Shakespeare and Bacon, Biometrika 62 (1975) 207–212.

[37] Hugh Craig, Arthur F. Kinney (Eds.), Shakespeare, Computers, and the Mystery of Authorship, Cambridge University Press, Cambridge, 2009.

[38] C. Labbé, D. Labbé, Inter-textual distance and authorship attribution Corneille and Molière, J. Quant. Ling. 8 (2001) 213–231.

[39] G.K. Zipf, Selected Studies of the Principle of Relative Frequency in Language, Harvard University Press, Cambridge, 1932.

[40] C. Manning, H. Schutze, Foundations of Statistical Natural Language Processing, MIT Press, Cambridge, 1999.

[41] J. Savoy, Machine Learning Methods for Stylometry: Authorship Attribution and Author Profiling, Springer International Publishing, Cham, 2020.

[42] L.Q. Ha, E.I. Sicilia-Garcia, J. Ming, F.J. Smith, Extension of Zipf's law to words and Phrases, in: COLING 2002: the 19th International Conference on Computational Linguistics, 2002, pp. 1–6.

[43] M.A. Montemurro, D.H. Zanette, New Perspectives on Zipf's law in linguistics: from single texts to large Corpora, Glottometrics 4 (2002) 87–99.

[44] I. Kanter, D.A. Kessler, Markov Processes: linguistics and Zipf's law, Phys. Rev. Lett. 74 (1995) 4559–4562.

[45] R. Harald Baayen, Analyzing Linguistic Data: A Practical Introduction to Statistics Using R, Cambridge University Press, Cambridge, 2008.

[46] J. Houvardas, E. Stamatatos, N-gram feature selection for authorship identification, in: J. Euzenat, J. Domingue (Eds.), *Artificial Intelligence: Methodology, Systems, and Applications*, Lecture Notes in Com- Puter Science, Springer, Berlin, Heidelberg, 2006, pp. 77–86.

[47] F. López-Escobedo, C.-F. Méndez-Cruz, G. Sierra, J. Solórzano-Soto, Analysis of stylometric Variables in long and short texts, Procedia - Social and Behavioral Sciences 95 (2013) 604–611.

[48] F. Mosteller, D.L. Wallace, Inference in an authorship problem, J. Am. Stat. Assoc. 58 (1963) 275–309.

[49] D.I. Holmes, R.S. Forsyth, The Federalist Revisited: New Directions in authorship attribution, Lit. Ling. Comput. 10 (1995) 111–127.

[50] S.K. Khamis, Review of inference and disputed authorship: the Federalist, Rev. Inst. Int. Stat./Rev. Int. Stat. Inst. 34 (1966) 277–279.

[51] M. Koppel, J. Schler, S. Argamon, Computational methods in authorship attribution, J. Am. Soc. Inf. Sci. Technol. 60 (2009) 9–26.

[52] M.B. Malyutov, Authorship attribution of texts: a review, Electron. Notes Discrete Math. 21 (2005) 353–357.

[53] B. Kjell, Authorship attribution of text samples using Neural networks and bayesian classifiers, Proc. IEEE Int. Conf. Syst. Man Cybern. 2 (1994) 1660–1664.

[54] G.K. Mikros, Blended authorship attribution: unmasking elena ferrante combining different author pro- filing methods, in: A. Tuzzi, M.A. Cortelazzo (Eds.), Drawing Elena Ferrante's Profile: Workshop Proceedings, Padova, 2018, pp. 85–95. Padova, UP.

[55] J. Burrows, Word-patterns and story-shapes: the statistical analysis of narrative style, Lit. Ling. Comput. 2 (1987) 61–70.

[56] E. Stamatatos, Ensemble-based author identification using character n-grams, Proceedings of the 3rd International Workshop on Text-based Information Retrieval 36 (2006) 41–46.

[57] W. Fucks, On mathematical analysis of style, Biometrika 39 (1952) 122–129.

[58] E. Stamatatos, On the robustness of authorship attribution based on character N-gram features symposium, J. Law Pol. 21 (2012) 421–440.

[59] F.J. Damerau, The use of function word frequencies as indicators of style, Comput. Humanit. 9 (1975) 271–280.

[60] J. C. Baker Pace, A test of authorship based on the rate at which new words enter an author's text, Lit. Ling. Comput. 3 (1988) 36–39.

[61] E.H. Simpson, Measurement of diversity, Nature 163 (1949) 688.

[62] P. Thoiron, Diversity index and entropy as measures of lexical richness, Comput. Humanit. 20 (1986) 197–202.

[63] G.U. Yule, The Statistical Study of Literary Vocabulary, 1944.

[64] G. Herdan, A new derivation and interpretation of Yule's 'characteristic'K, Zeitschrift für angewandte Mathematik und Physik ZAMP 6 (1955) 332–339.

[65] P.E. Bennett, The statistical measurement of a stylistic trait in julius caesar and as you like it, Shakespeare Q. 8 (1957) 33–50.

[66] H.S. Sichel, On a distribution law for word frequencies, J. Am. Stat. Assoc. 70 (1975) 542–547.

[67] D.A. Ratkowsky, L. Hantrais, Tables for comparing the richness and structure of vocabulary in texts of different lengths, Comput. Humanit. 9 (2) (1975) 69–75.

[68] A. Q. Morton Once, A test of authorship based on words which are not repeated in the sample, Lit. Ling. Comput. 1 (1986) 1–8.

[69] M.W.A. Smith, Hapax legomena in prescribed positions: an investigation of recent proposals to resolve problems of authorship, Lit. Ling. Comput. 2 (1987) 145–152.

[70] B. Mandelbrot, A note on a class of skew distribution functions: analysis and critique of a paper by H. A. Simon, Inf. Control 2 (1959) 90–99.

[71] J. Savoy, Authorship attribution based on specific vocabulary, ACM Trans. Inf. Syst. 30 (12) (2012) 1–12:30.

[72] F. Peng, D. Schuurmans, S. Wang, V. Keselj, Language independent authorship attribution using charac- ter level Language models, in: Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics - Volume 1, Association for Computational Lin- guistics, USA, 2003, pp. 267–274.

[73] C. Akimushkin, D.R. Amancio, O.N. Oliveira, On the role of words in the Network structure of texts: application to authorship attribution, Phys. Stat. Mech. Appl. 495 (2018) 49–58.

[74] E. Stamatatos, Intrinsic plagiarism detection using character n-gram profiles, in: Proceedings of the SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse, 2009, pp. 38–46. *PAN 2009)*.

[75] J.F. Hoorn, S.L. Frank, W. Kowalczyk, F. van der Ham, Neural Network identification of poets using letter sequences, Lit. Ling. Comput. 14 (1999) 311–338.

[76] G. Frantzeskou, E. Stamatatos, S. Gritzalis, S. Katsikas, Effective identification of source code authors using byte-level information, in: *Proceedings Of the 28th International Conference On Software Engineering*, ICSE '06, Association for Computing Machinery, New York, NY, USA, 2006, pp. 893–896.

[77] H. Baayen, H. van Halteren, A. Neijt, F. Tweedie, An experiment in authorship attribution, 6th JADT 1 (2002) 69–75.

[78] P. Juola, H. Baayen, A controlled-corpus experiment in authorship identification by cross-entropy, in: Literary and Linguistic Computing, 2003.

[79] J. Savoy, Is starnone really the author behind ferrante? Digital Scholarship in the Humanities 33 (2018) 902–918.

[80] E. Stamatatos, Author identification using imbalanced and limited training texts, in: 18th International Workshop on Database and Expert Systems Applications (DEXA 2007), 2007, pp. 237–241.

[81] Ö. Uzuner, B. Katz, A comparative study of language models for book and author recognition, in: R.t Dale, K.-F. Wong, J. Su, O.Y. Kwong (Eds.), *Natural Language Processing – IJCNLP 2005*, Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2005, pp. 969–980.

[82] R. Zheng, J. Li, H. Chen, Z. Huang, A framework for authorship identification of online messages: writing-style features and classification techniques, J. Am. Soc. Inf. Sci. Technol. 57 (2006) 378–393.

[83] H. Wu, Z. Zhang, Q. Wu, Exploring syntactic and semantic features for authorship attribution, Appl. Soft Comput. 111 (2021) 107815.

[84] U. Athira, S.M. Thampi, An author-specific-model-based authorship analysis using psycholinguistic aspects and style word patterns, J. Intell. Fuzzy Syst. 34 (2018) 1453–1466.

[85] R.S. Forsyth, D.I. Holmes, Feature-finding for text classification, Lit. Ling. Comput. 11 (1996) 163–174.

[86] M.L. Jockers, D.M. Witten, A comparative study of machine learning methods for authorship attribu- tion, Lit. Ling. Comput. 25 (2) (2010) 215–223.

[87] P. Juola, Verifying authorship for forensic purposes: a computational protocol and its validation, Forensic Sci. Int. 325 (2021) 110824.

[88] J. Savoy, Authorship attribution based on a probabilistic topic model, Inf. Process. Manag. 49 (2013) 341–354.

[89] A.S. Altheneyan, M.E.B. Menai, Näive Bayes classifiers for authorship attribution of Arabic texts, Journal of King Saud University-Computer and Information Sciences 26 (2014) 473–484.

[90] C.S. Brinegar, Mark twain and the quintus curtius snodgrass letters: a statistical test of authorship, J. Am. Stat. Assoc. 58 (1963) 85–96.

[91] E. Elhaik, Principal component analyses (PCA)-Based findings in population genetic studies are highly biased and must Be reevaluated, Sci. Rep. 12 (2022) 14683.

[92] P.M. Kroonenberg, M.J. Greenacre, Correspondence analysis, in: Encyclopedia of Statistical Sciences, John Wiley & Sons, Ltd, 2006.

[93] V. Koshti, Cumulative sum control chart, Int. J. Phys. Math. Sci. 1 (2011) 28–32. ISSN: 2277–2111.

[94] R.A. Hardcastle, Forensic linguistics: an assessment of the CUSUM method for the determination of authorship, J. Forensic Sci. Soc. 33 (1993) 95–106.

[95] B. Kjell, Authorship determination using letter pair frequency features with neural Network classifiers, Lit. Ling. Comput. 9 (1994) 119–124.

[96] D. Lowe, R. Matthews, Shakespeare vs. Fletcher: a stylometric analysis by radial basis functions, Comput. Humanit. 29 (1995) 449–461.

[97] F.J. Tweedie, S. Singh, D.I. Holmes, Neural Network applications in stylometry: the federalist papers, Comput. Humanit. 30 (1996) 1–10.

[98] J. Diederich, J. Kindermann, E. Leopold, G. Paass, Authorship attribution with support vector machines, Appl. Intell. 19 (2003) 109–123.

[99] J. Pearl, The limitations of opaque learning machines, in: J. Brockman (Ed.), Possible Minds: Twenty-Five Ways of Looking at AI, Penguin Press, 2019, pp. 13–19.

[100] S. Bozza, F. Taroni, R. Marquis, M. Schmittbuhl, Probabilistic evaluation of handwriting evidence: likelihood ratio for authorship, J. Roy. Stat. Soc.: Series C (Applied Statistics) 57 (2008) 329–341.

[101] S. Bozza, J. Broséus, P. Esseiva, F. Taroni, Bayesian classification criterion for forensic multivariate data, Forensic Sci. Int. 244 (2014) 295–301.

[102] A. Biedermann, The strange persistence of (source) "identification" claims in forensic literature through descriptivism, diagnosticism and machinism, Forensic Sci. Int.: Synergy 4 (2022) 100222.

[103] F. Taroni, S. Bozza, A. Biedermann, Decision theory, in: D.L. Banks, K. Kafadar, D. H. Kaye, M. Tackett (Eds.), Handbook of Forensic Statistics, Chapman and Hall/ CRC, Boca Raton, 2020, pp. 103–130.

[104] F. Taroni, S. Bozza, A. Biedermann, The logic of inference and decision for scientific evidence, in: C. Dahlman, A. Stein, G. Tuzet (Eds.), Philosophical Foundations of Evidence Law, Oxford University Press, 2021, pp. 251–266.

[105] A. Biedermann, Machine Learning Enthusiasts Should Stick to the Facts. Response to Morrison et al. (2022), Forensic Sci. Int.: Synergy 4 (2022) 100229.

[106] G. Jackson, The scientist and the scale of justice, Sci. Justice 40 (2000) 81–85.

[107] A. Tuzzi, M.A. Cortelazzo, What is elena ferrante? A comparative analysis of a secretive bestselling Italian writer, Digital Scholarship in the Humanities 33 (2018) 685–702.

[108] M. Kocher, J. Savoy, Author clustering using SPATIUM, in: 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL), 2017, pp. 1–4.

[109] European Network of Forensic Science Institutes (ENFSI), ENFSI Guideline for Evaluate Reporting in Foren- Sic Science, 2016.

[110] F. Taroni, P. Garbolino, S. Bozza, C. Aitken, The Bayes' factor: the coherent measure for hypothesis confirmation, Law Probab. Risk 20 (1) (2021) 15–36.

[111] S. Bozza, V. Cammarota, C.-A. Roten, V. Roten, A. Jover, F. Taroni, The Bayes Factor to Assess the Value of Evidence for Stylometric Data: a Study on Molière's Comedies, School of Criminal Justice - The University of Lausanne, 2024. Technical report.

[112] E. Stamatatos, N. Fakotakis, G. Kokkinakis, Automatic text categorization in terms of genre and author, Comput. Ling. 26 (2000) 471–495.

[113] Y. Zhao, J. Zobel, Effective and scalable authorship attribution using function words, in: Asia Information Retrieval Symposium, Springer, 2005, pp. 174–189.

[114] M. Yang, K.-P. Chow, Authorship attribution for forensic investigation with thousands of authors, in: Nora Cuppens-Boulahia, Frédéric Cuppens, Sushil Jajodia, Anas Abou El Kalam, Thierry Sans (Eds.), *ICT Systems Security And Privacy Protection*, IFIP Advances in Information and Communication Technology, Springer, Berlin, Heidelberg, 2014, pp. 339–350.

[115] P. Juola, Authorship studies and the dark side of social media analytics, J. Univers. Comput. Sci. 26 (2020) 156–170.

[116] P. Juola, G.K. Mikros, Cross-linguistic stylometric features: a preliminary investigation, in: International Conference on Statistical Analysis of Textual Data, Nice, France, 2016.

[117] P. Juola, Detecting contract cheating via stylometric methods, in: Proceedings on the Conference on Plagia- Rism across Europe and beyond, 2017, pp. 187–198.

[118] M.W.A. Smith, Forensic stylometry: a theoretical basis for further developments of practical methods, J. Forensic Sci. Soc. 29 (1989) 15–33.

[119] D.C. Ison, Detection of online contract cheating through stylometry: a pilot study, Online Learn. 24 (2020) 142–165.

[120] I. Rosengren, Style as choice and deviation, Style 6 (1972) 3–18.

[121] S. Evert, T. Proisl, F. Jannidis, I. Reger, S. Pielström, C. Schöch, T. Vitt, Understanding and explaining delta measures for authorship attribution, Digital Scholarship in the Humanities 32 (2017) 4–16.

[122] G.R. McMenamin, Declaration of Gerald McMenamin, Ceglia v. Zuckerberg., 2011.

[123] M. Eder, Elena ferrante: a virtual author, in: A. Tuzzi, M.A. Cortelazzo (Eds.), Drawing Elena Ferrante's Profile: Workshop Proceedings vol. 31, 2018. Padova UP.

[124] H.M. Gómez-Adorno, G. Rios-Toledo, J.-P. Posadas-Durań, G. Sidorov, G. Sierra, Stylometry-based approach for detecting writing style changes in literary texts, Computación y Sistemas 22 (2018).

[125] J. Savoy, Text Categorization with Style (And Class), 2017.

[126] R. Marquis, Etude de caractères manuscrits : de la caractérisation morphologique à l'individualisation du scripteur. PhD thesis, Université de Lausanne, Faculté de droit, des sciences criminelles et d'administration publique, 2007.

[127] M. Kestemont, K. Luyckx, W. Daelemans, T. Crombez, Evaluating unmasking for cross-genre Au- thorship verification, in: J.C. Meister (Ed.), 7th Annual International Conference of the Alliance of Digital Humanities Organizations, Hamburg University Press, 2012, pp. 249–251.

[128] C.E. Chaski, Empirical evaluations of language-based author identification techniques, Int. J. Speech Lang. Law 8 (2001) 1–65.

[129] Maciej Eder, Does size matter? Authorship attribution, small samples, big problem, Lit. Ling. Comput. 30 (2) (2015) 167–182. ISSN 0268-1145.

[130] Luyckx Kim, Walter Daelemans, Authorship attribution and verification with many authors and limited data, in: Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), 2008, pp. 513–520. Manchester, UK. Coling 2008 Organizing Committee.

[131] X. Li, The study on forensic linguistic analysis of applied linguistics based on data analysis, in: S. Lin, X. Huang (Eds.), *Advances In Computer Science, Environment, Ecoinformatics, and Education*, Commu- Nications in Computer and Information Science, Springer, Berlin, Heidelberg, 2011, pp. 528–532.

[132] S.E. Fienberg, S.H. Krislov, M.L. Straf, Understanding and evaluating statistical evidence in litigation, Jurimetrics 36 (1995) 1–32.

[133] J. Gonzalez-Rodriguez, P. Rose, D. Ramos, D.T. Toledano, J. Ortega-Garcia, Emulating DNA: rigorous quantification of evidential weight in transparent and testable forensic speaker recognition, IEEE Trans. Audio Speech Lang. Process. 15 (2007) 2104–2115.

[134] M. Redmayne, Science, evidence and logic, Mod. L. Rev. 59 (1996) 747–760.

[135] S. Ishihara, Strength of linguistic text evidence: a fused forensic text comparison system, Forensic Sci. Int. 278 (2017) 184–197.

[136] S. Bozza, C.-A. Roten, A. Jover, V. Cammarota, L. Pousaz, F. Taroni, A model-independent redundancy measure for human versus ChatGPT authorship discrimination using a Bayesian probabilistic approach, Nature Scientific Reports 13 (1) (2023) 19217.

[137] C. Champod, D. Baldwin, F. Taroni, J. Buckleton, Firearm and tool marks identification: the bayesian approach, AFTE journal 35 (3) (2003) 307–316.

[138] F. Riva, Etude sur la valeur indicielle des traces présentes sur les douilles, 2011.

[139] J. Gonzalez-Rodriguez, J. Fiérrez-Aguilar, J. Ortega-Garcia, Forensic identification reporting using Au- tomatic speaker recognition systems, in: 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03) vol. 2, IEEE, 2003, pp. II–93.

[140] F. Taroni, S. Bozza, C. Aitken, Decision analysis in forensic science, J. Forensic Sci. 50 (2005) 894–905.