Language and Law

Linguagem e Direito

VOLUME 5.2

ISSN 2183-3745

Language and Law
Linguagem e Direito

ISSN: 2183-3745 (online)
Volume 5, Issue 2, 2018

Publisher / Editora
Faculdade de Letras da Universidade do Porto

U.PORTO Faculdade de Letras Universidade do Porto
Centro de Linguística da Universidade do Porto
Linguística Forense
ALIDI ASSOCIAÇÃO DE LINGUAGEM & DIREITO

Language and Law / Linguagem e Direito

Language and Law / Linguagem e Direito is a free, exclusively online peer-reviewed journal published twice a year. It is available on the website of the Faculty of Arts of the University of Porto, at http://ojs.letras.up.pt/index.php/LLLD/ .

All articles should be submitted by email to the journal email address (llldjournal@gmail.com) or via the system. See the guidelines for submission at the end of this issue.

Requests for book reviews should be sent to llldjournal@gmail.com.

Language and Law / Linguagem e Direito é uma revista gratuita publicada exclusivamente online, sujeita a revisão por pares, publicada semestralmente e disponível no website da Faculdade de Letras da Universidade do Porto, em http://ojs.letras.up.pt/index.php/LLLD/ .

Os materiais para publicação deverão ser enviados por email para o endereço da revista (llldjournal@gmail.com) ou através do sistema, e devem seguir as instruções disponíveis no final deste volume.

As propostas de recensão de livros devem ser enviadas para llldjournal@gmail.com.

Indexing and abstracting / Indexação e bases de dados bibliográficas
Language and Law / Linguagem e Direito is covered by the following abstracting and indexing service:

A Language and Law / Linguagem e Direito encontra-se indexada e catalogada na seguinte bases de dados:

ERIH PLUS: European Reference Index for the Humanities and the Social Sciences
Google Scholar
Sherpa Romeo
Jornals for Free
CrossRef
Portal RCAAP
QUALIS Periódicos
JURN

PUBLISHED BIANNUALLY ONLINE / PUBLICAÇÃO SEMESTRAL ONLINE

ISSN: 2183-3745

Contents / Índice

Introduction

Rui Sousa-Silva & Malcolm Coulthard

University of Porto, Portugal & Federal University of Santa Catarina, Brazil

Language and Law/Linguagem e Direito is delighted to publish this Special Issue devoted to Computational Forensic Linguistics. It contains a balanced collection of six articles by both forensic linguists and computer scientists/computational linguists which are exclusively dedicated to the area. We think the issue will come to be seen as a significant addition to the body of research into forensic linguistics and so we are especially pleased that, as with all articles published in *Language and Law/Linguagem e Direito*, readers will have free access – there is no paywall and no cost for authors!

The work of forensic linguists is nowadays inescapably linked to Computational Forensic Linguistics. Whether the forensic linguistic task consists of analysing the authorship of a questioned document, detecting plagiarism, analysing the (disputed) meaning of a text or text excerpt, investigating courtroom or police discourse, or even translating and interpreting in forensic contexts, a competent application of computational tools and techniques is crucial. Indeed, over the last decades not only has the nature of forensic applications evolved dramatically, but so has the volume of text needing to be analysed increased exponentially. Additionally, thanks to more recent technological developments, a significant proportion of criminal activity has started taking place online, so methods used in the past have had to be constantly updated to handle the new challenges. Computational forensic linguistics is ideally placed to assist forensic linguists address these challenges.

This Special Issue opens with 'Computational Forensic Authorship Analysis: Promises and Pitfalls'. In this article, Shlomo E. Argamon surveys from a practitioner's perspective the different types of computational authorship analysis methods and their components, with a view to ensuring reliability. The author identifies and discusses specifically some of the pitfalls potentially faced by an analyst when applying the methodology, and eventually offers guidance to practitioners.

The issue continues with Andrea Nini's article 'Developing forensic authorship profiling', which approaches authorship profiling in a forensic context. As the author argues, current methods lack the transparency offered by certain computational

techniques, and so fail to meet the standards required for forensic applications. The article reports an experiment he conducted to show how previously established findings related to stylistic variation in English for gender, age and social class also apply to forensic texts. The author concludes by demonstrating the relevance of linguistically-motivated research into forensic authorship profiling.

The volume continues with Sheila Queralt's article 'The creation of Base Rate Knowledge of linguistic variables and the implementation of likelihood ratios to authorship attribution in forensic text comparison', in which she explores the issue of reliability in forensic authorship comparison. In order to guarantee reliability that is comparable with other forensic disciplines, the author proposes the implementation of statistical techniques and argues that such a method assists, not only the courts, but also the linguistic experts.

Patrick Juola then approaches the topic of professionalisation of forensic science through the development of standards and protocols. In his article, entitled 'The Rowling Protocol, Steven Bannon, and Rogue POTUS Staff: a Study in Computational Authorship Attribution', the author applies a systematic protocol for authorship verification (previously used in his analysis of the Rowling case) to another high-profile case: the "Rogue POTUS Staff" (self-described as "The unofficial resistance team inside the White House. We pull back the curtain to expose the real workings inside this disastrous, frightening Administration.").

The next article, 'On the Implications of the General Data Protection Regulation on the Organisation of Evaluation Tasks', by Francisco Rangel and Paolo Rosso, focuses on the challenges that the General Data Protection Regulation (GDPR) of the European Union presents to the organisation of evaluation tasks. As these tasks, which are frequently hosted as part of computational linguistics conferences to test the performance of different computer systems, involve collecting and making available large volumes of data collected from the Internet, in general, and from social media platforms, in particular, they must now meet the stringent requirements of the GDPR. The authors build upon experience gained from the organisation of tasks such as PAN to discuss especially how the collection and distribution of the data used in those tasks comply or fail to comply with European regulations. They propose a methodology to follow when organising such tasks and conclude with a discussion of a practical case.

The volume ends with the article 'Computational Forensic Linguistics: An Overview of Computational Applications in Forensic Contexts', by Rui Sousa-Silva which reviews a significant body of the available literature on computational linguistics approaches that are (or can potentially be) used in forensic linguistics applications. Such applications include authorship analysis, authorship profiling and stylometry, plagiarism detection and analysis and cybercrime, as well as less high-profile applications such as meaning analysis. The article concludes with a discussion of both the potential and the limitations of computational approaches to forensic linguistic analysis, and the future implications for forensic linguistics.

We hope that this collection of articles gives readers an insight into the exciting field of computational forensic linguistics and encourages all those who share an interest in the area to follow one of these lines of research. Others should find in the research presented reasons for employing computational applications in forensic linguistics

casework. Finally, computer scientists (and computational linguists) will hopefully gain a deeper understanding of the challenges driving forensic linguistics research.

We hope you find reading this special issue a rewarding experience – it's been a pleasure editing it!

Rui Sousa-Silva
University of Porto
Portugal

Malcolm Coulthard
Federal University of Santa Catarina
Brazil

Nota Introdutória

Rui Sousa-Silva & Malcolm Coulthard

Universidade do Porto, Portugal & Universidade Federal de Santa Catarina, Brasil

A *Language and Law/Linguagem e Direito* tem o prazer de publicar este número especial, dedicado à Linguística Forense Computacional. O número inclui seis artigos originais nesta área redigidos, quer por linguistas forenses, quer por linguistas computacionais/especialistas em ciências dos computadores, oferecendo, assim, um conjunto de artigos muito equilibrado. Acreditamos que este número representará um contributo significativo para o volume de publicações e investigação em linguística forense; por isso, é, para nós, um prazer enorme que, tal como acontece com todos os artigos publicados na *Language and Law/Linguagem e Direito*, os leitores tenham acesso gratuito a este número – não existe acesso pago –, tal como não existem custos para os autores!

O trabalho dos linguistas forenses encontra-se, atualmente, inevitavelmente associado à Linguística Forense Computacional. Quer se trate de analisar a autoria de um documento suspeito, quer se trate de detetar plágio, analisar o significado (questionado) de um texto ou expressão, investigar o discurso da sala de audiências ou da polícia, ou, inclusivamente, desempenhar tarefas de tradução ou interpretação em contextos forenses, a utilização competente de técnicas e ferramentas computacionais é essencial para o desempenho das tarefas de linguista forense. De facto, ao longo das últimas décadas assistimos, não só a uma evolução significativa da natureza das aplicações forenses, como também a um aumento exponencial do volume de texto para análise. Além disso, graças aos mais recentes desenvolvimentos tecnológicos, uma parte significativa da atividade criminal passou a decorrer online, pelo que os métodos tradicionalmente utilizados têm de ser constantemente atualizados, de modo a poderem dar resposta aos novos desafios. A linguística forense computacional oferece as condições ideais para ajudar os linguistas forenses a darem essas respostas.

Este número especial abre com 'Computational Forensic Authorship Analysis: Promises and Pitfalls', em que Shlomo E. Argamon faz uma revisão dos diferentes tipos de métodos de análise de autoria computacional e dos seus componentes na perspetiva de um profissional, de modo a assegurar a sua fiabilidade. O autor identifica e discute especificamente algumas das armadilhas com as quais os/as analistas se podem

confrontar na aplicação da metodologia, e fornece algumas orientações destinadas a profissionais.

O número continua com o artigo de Andrea Nini 'Developing forensic authorship profiling', que aborda a determinação de perfis de autoria em contextos forenses. Como defende o autor, aos métodos utilizados atualmente falta a transparência de determinadas técnicas computacionais, o que não lhes permite satisfazer os requisitos exigidos pelas aplicações forenses. O artigo descreve uma experiência realizada pelo autor que revela de que modo resultados de estudos efetuados anteriormente relacionados com variação estilística em inglês relativos às categorias de género, idade e classe social também são aplicáveis a textos forenses. O autor termina demonstrando a relevância da investigação em perfis de autoria forense motivada linguisticamente.

O volume prossegue com o artigo de Sheila Queralt 'The creation of Base Rate Knowledge of linguistic variables and the implementation of likelihood ratios to authorship attribution in forensic text comparison', no qual explora a questão da fiabilidade em comparação de autoria forense. De modo a garantir uma fiabilidade comparável com outras disciplinas forenses, a autora propõe a implementação de técnicas estatísticas e defende que um método deste tipo auxilia, não só os tribunais, mas também os/as peritos/as linguísticos/as.

Patrick Juola aborda, então, o tema da profissionalização da ciência forense através do desenvolvimento de normas e protocolos. No seu artigo, entitulado 'The Rowling Protocol, Steven Bannon, and Rogue POTUS Staff: a Study in Computational Authorship Attribution', o autor aplica um protocolo sistemático de verificação de autoria (utilizado anteriormente na sua análise do caso Rowling) a outro caso mediático: o "Rogue POTUS Staff" (que se auto-descreve como "A equipa de resistência não-oficial no interior da Casa Branca. Puxamos as cortinas para expor as verdadeiras operações no seio desta Administração desastrosa e assustadora."), oferecendo algumas conclusões sobre a análise.

O artigo seguinte, 'On the Implications of the General Data Protection Regulation on the Organisation of Evaluation Tasks', da autoria de Francisco Rangel e Paolo Rosso, centra-se nos desafios que o Regulamento Geral de Proteção de Dados (RGPD) da União Europeia representa para a organização de tarefas de avaliação. Uma vez que estas tarefas, normalmente integradas em congressos de linguística computacional com vista a testar o desempenho de diferentes sistemas informáticos perante o mesmo problema, implicam a recolha e disponibilização de grandes quantidades de dados recolhidos da Internet, em geral, e de redes sociais, em particular, é agora obrigatório que cumpram os exigentes requisitos do RGPD. Os autores baseiam-se na experiência decorrente da organização de competições como o PAN para discutir em particular de que modo a recolha e distribuição dos dados utilizados nessas tarefas cumprem ou, pelo contrário, infringem as regulamentações europeias. Os autores propõem uma metodologia a seguir para organizar tarefas deste tipo, terminando o artigo com a discussão de um caso prático.

O número termina com o artigo 'Computational Forensic Linguistics: An Overview of Computational Applications in Forensic Contexts', de Rui Sousa-Silva, que apresenta uma revisão de um número significativo de referências bibliográficas sobre abordagens de linguística computacional que são (ou podem potencialmente ser) utilizadas em

aplicações de linguística forense. Entre estas, incluem-se a análise de autoria, perfis de autoria e estilometria, análise e deteção de plágio e cibercrime, bem como aplicações menos conhecidas, como análise de significados. O artigo termina com uma discussão do potencial e das limitações das abordagens computacionais à análise linguística forense, bem como das implicações futuras para a linguística forense.

Esperamos que este conjunto de artigos proporcione aos leitores uma perspetiva aprofundada sobre a empolgante área da linguística forense computacional e incentive todos aqueles com interesse nesta área a seguir uma destas linhas de investigação. Outros deverão encontrar na investigação apresentada motivos suficientes para utilizar aplicações computacionais em casos práticos de linguística forense. Finalmente, os especialistas em ciências dos computadores (e os linguistas computacionais) irão, esperamos, ficar a conhecer mais pormenoridamente os desafios por detrás da investigação em linguística forense.

Esperamos que a leitura deste número especial lhe proporcione uma experiência gratificante; para nós, editores da revista, foi um prazer editá-lo!

Rui Sousa-Silva
Universidade do Porto
Portugal

Malcolm Coulthard
Universidade Federal de Santa Catarina
Brasil

Computational Forensic Authorship Analysis: Promises and Pitfalls

Shlomo Engelson Argamon

Illinois Institute of Technology, USA

Abstract. *The authorship of questioned documents often constitutes important evidence in criminal and civil cases. Linguistic stylistic analysis can often help to determine authorship. Computational methods have been applied to authorship analysis in academia for decades, and in recent years have achieved the levels of reliability needed for application to real-world cases. This article surveys the different types of computational authorship analysis methods and their components in a practical vein—describing the assumptions each makes, the analytic controls they require, and the tests needed to measure and ensure their reliability. Specifically, I discuss many of the potential pitfalls in their application, to guide practitioners in more effectively achieving trustworthy and understandable results. It must always be remembered, though, that there is no substitute for expertise, experience, and careful human judgment.*

Keywords: *Authorship, computational forensic linguistics, computational autorship analysis, reliability.*

Resumo. *A autoria de documentos questionados constitui, muitas vezes, prova importante em casos civis e criminais. A análise linguística estilística ajuda frequentemente a determinar a autoria. Na academia, há várias décadas que os métodos computacionais são aplicados à análise de autoria, tendo, recentemente, alcançado os níveis de fiabilidade necessários para aplicação em casos reais. Este artigo apresenta uma revisão dos diversos tipos de métodos de análise de autoria computacional e os seus diversos componentes numa perspetiva prática—descrevendo os pressupostos de cada um, os controlos analíticos de que necessitam, e os testes necessários para medir e assegurar a sua fiabilidade. Especificamente, discuto muitas das possíveis armadilhas inerentes à sua aplicação, de modo a ajudar os peritos fornecendo-lhes orientações para alcançarem resultados mais fiáveis e compreensíveis. Não podemos esquecer, contudo, que não existe qualquer substituto para a especialização, experiência e cuidadoso julgamento humano.*

Palavras-chave: *Autoria, linguística forense computacional, análise de autoria computacional, fiabilidade.*

Introduction

Computational methods for authorship attribution have grown in importance for forensics as they have become more accurate and more applicable to real-world situations. A well-publicized recent case of computational authorship attribution (if not in a forensic context) was the 2013 computational unmasking of J. K. Rowling as the author of the novel *The Cuckoo's Calling* by (independently) Peter Millican and Patrick Juola (Mostrous, 2013; Zimmer, 2013). They were contacted by London's *Sunday Times* to confirm a tip that Rowling had pseudonymously written the book. The two researchers independently performed computational stylometric analyses that pointed towards Rowling as a more likely author than some other plausible candidates; when shown the evidence, she reluctantly admitted that she was the author.

Of course, it is rare for a forensic authorship question to end with an unequivocal confession, and so the question of the strength and reliability of the evidence adduced is critical. *Daubert's* criterion that a method have "known or potential rate of error" is not a simple question to answer, since performance of any method will depend greatly on the specifics of the case. It can be tricky to ensure that the right analysis method is used for the task, to design the analysis protocol to produce reliable results, and to properly assess the strength of the resulting evidence. There are many parameters that must be determined and set, and there are no simple formulas for doing so that are valid in all cases. Always expert judgment is a key factor.

This article provides guidelines for using computational authorship attribution in the forensic context (and for critiquing such use). Specifically, my aims here are to show (i) how current computational methods can be used for authorship attribution, (ii) the promise they bring to forensic authorship analysis as a complement to traditional linguistic techniques, and (iii) how to recognize and avoid common methodological pitfalls in their application.

Overview of the Process

The process of applying computational authorship attribution starts with three key choices (partly externally constrained):

- Choose an attribution algorithm/method to use;
- Create a corpus comprising two or three subcorpora: the questioned texts (Q) of unknown authorship (provided as part of the case), a set of known texts (K) by candidate authors (usually provided by the attorneys), and possibly (depending on the method to be used) some comparison texts (C);
- Determine what features of the texts to extract and what measurements of their occurrence to use to characterize each text.

While these choices can be made separately—different methods can be applied to the same corpora and features, different feature sets can be used with a single method, etc.—they are significantly interrelated. Some features may work better with some methods than others, and the choice of method may have strong implications for how the corpus is constructed and vice versa.

Given a corpus, method, and features, the attribution process is as follows, in broad outline:

1. Evaluate the chosen method on texts of known authorship to establish the reliability of the method for the given case;

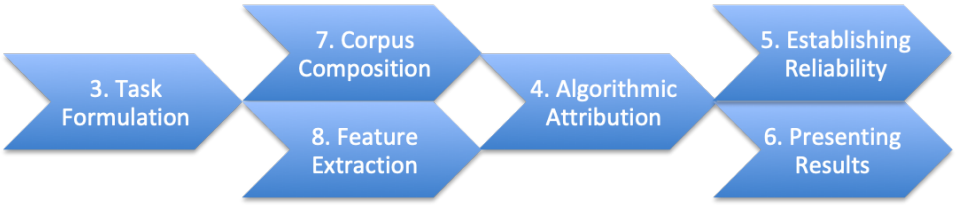

Figure 1. The flow of the overall process of authorship attribution, showing which sections of this article deal with each subtask.

2. Apply the chosen method to K, Q, and C to form an analysis of the authorship of the texts in Q;
3. Evaluate the meaning and significance of attribution results in the context of the given case.

There are different ways to implement each step of the process, some of them valid only for certain methods or in certain circumstances; we will discuss these considerations below.

In the remainder of the article I will discuss the considerations that should go into these six elements of the process, and the pitfalls that must be avoided to ensure trustworthy results. Each of the following sections treat one main aspect of the process and method of computational authorship attribution, as depicted in Figure 1. The article does not follow the order of the process, for expository reasons.

It must be emphasized that this article is naturally only a roadmap, and the mere fact that an analysis avoids the pitfalls discussed herein cannot guarantee its validity—expert judgment must always be applied to the specifics of any case.

Task Formulation

Before discussing different computational authorship attribution methods, we must first discuss the variety of attribution tasks that can be addressed. Different formulations of the task will be appropriate for different cases, as we will see.

The simplest task formulation, *author classification* is where a set of specific candidate authors with known writings is given. For example, the *Federalist Papers* are a series of articles published pseudonymously by Alexander Hamilton, James Madison, and John Jay in 1787 and 1788 to promote the ratification of the new United States Constitution. In this case, famously addressed in Mosteller and Wallace's (1964) landmark stylometric study, there are three candidate authors, and the problem is to classify each article to its correct author. Or consider the authorship question of the various sections of the late 16th Century play *The Raigne of King Edward the Third*, whose authorship is widely disputed. Much of the play is attributed to Shakespeare, but many sections are variously attributed to several other period playwrights, mainly Thomas Kyd, Christopher Marlowe, Michael Drayton, and George Peele. The attribution question for a particular section (say, one scene) could be formulated as "Which of these five individuals wrote this section?"

In general, the larger the number of candidates, the harder the task is to solve. Even if a set of candidate authors is known, it is often necessary to consider the possibility that some unknown author outside that set is the actual author (i.e., to allow "unknown" as an answer to the classification question). This setting, *open-set attribution*, is more difficult

to solve reliably than *closed-set* attribution, where the candidate set is known (or can be assumed) to contain all possible authors of the questioned text.

An important form of open-set attribution is *author verification*, where there is only one candidate author A and the task is to determine whether or not that individual was the author of the questioned document or not (Koppel *et al.*, 2007; Halteren, 2007; Koppel *et al.*, 2007). One important version of verification is when we are asked whether two documents X and Y were authored by the same person (Koppel *et al.*, 2012b).

As an example of verification, consider the question of whether the book *The Cuckoo's Calling* was written by J. K. Rowling, or not, as mentioned above. To analyze this question, Patrick Juola compared the style of the book with that of one other book by each of J. K. Rowling and three other British mystery authors, Ruth Rendell, P. D. James, and Val McDermid. The question was whether Rowling was a noticeably more likely author than the other three, which would provide some evidence for or against her authorship of *The Cuckoo's Calling*.

A solution to verification can also be used to solve general open-set author classification by comparing Q to the known documents for each candidate and attributing it to the author whose documents are most reliably same-authored with Q, and if none are, giving the result "unknown".

Verification is more difficult than classification, and requires different methods, since the alternatives include everyone in the world other than A.

In cases where a specific set of candidate authors is not available, *authorship profiling* can sometimes be useful, determining demographic and social characteristics of the author based on language use. Such profiling is based on comparing features of Q with features drawn from analysis of large datasets labeled for the profile categories of interest, such as author age, sex, education, linguistic background, and the like. As a general rule, due to its broader conclusions, authorship profiling is more useful for investigations rather than for evidence of specific authorship.

Pitfall 0(a) (Match the task formulation to the case) *Different formulations of authorship attribution make different assumptions about the nature of the data and the question to be answered. Make sure that your formulation of the problem matches the structure and evidential requirements of the case.*

Algorithmic Attribution

Methodological foundations

Most of the methods for solving the above problems rely on a fundamental notion of computing form of *stylometric similarity* and comparing its values for different texts.

There are many ways to devise a similarity measure for this purpose, and we will discuss some of the details of doing so below. In the vast majority of approaches, a similarity measure is constructed by first identifying a number of textual features which are presumed to be more-or-less indicative of style and authorship. The collection of the frequencies of these features in a given text then is considered to characterize the style of the text. For example, in one of Mosteller and Wallace's (1964) foundational studies of the *Federalist Papers*, they used a set of 68 function words as features. They thus characterized each of the Federalist Papers by a vector of 68 numbers, each the frequency in

the document of one of the function words from their list. In Section 'Feature Extraction' below we discuss the choice of features and how this may affect the reliability of authorship attribution.

Given a set of features, measuring the similarity of two texts comes down to measuring the similarity between two numeric vectors representing the frequencies of all the features in each of the documents. The more similar are the corresponding frequencies, the more similar the two texts are, in terms of the features that have been counted. A number of different mathematical formulations have been proposed for calculating a score for measuring similarity—the most commonly used today are:

- *cosine similarity*, commonly used in information retrieval (Salton and Lesk, 1968), computed for two vectors $\langle x_1, x_2, \cdots, x_n \rangle$ and $\langle y_1, y_2, \cdots, y_n \rangle$ as

$$\frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2 \sum_i y_i^2}}$$

- *min-max similarity* (or *Ruzicka similarity*), which has been recently shown to be particularly effective in authorship attribution applications (Kestemont *et al.*, 2016; Halvani *et al.*, 2018), computed as

$$\frac{\sum_i \min(x_i, y_i)}{\sum_i \max(x_i, y_i)}$$

This *feature-vector approach* to computing a measure of stylometric similarity between two texts comprises the steps of:

1. Identify the features of interest in each text;
2. Count the number of occurrences of each feature type in the texts, and normalize them to compute relative frequencies (as a fraction of total tokens in each respective text), giving a numeric feature vector for each text;
3. Compute a similarity score from the two vectors.

The precise character of the resulting similarity measure will depend on what textual features are chosen, how frequencies are normalized, and what similarity scoring function is used. All of these must be taken into account when comparing different methods.

Other document representations have also been used to construct useful similarity measures for authorship attribution. Similarity of graph representations of word type collocations (Arun *et al.*, 2009; Vilariño *et al.*, 2013) in documents can be compared by measuring the similarity of the graphs directly. Sequence-based "string kernel" methods (Lodhi *et al.*, 2001; Cancedda *et al.*, 2003; Xing *et al.*, 2010), developed for general text and genome comparison can also be used. In each case, the correlation of the chosen similarity measure with likelihood of authorship (and, as far as possible, independence of topic and text type) for the relevant texts must be established.

Stylometry and attribution

Now, let us suppose that we have in hand a reliable stylometric similarity measure M, such that we can assume that the likelihood that two texts have the same author is (roughly) proportional to the similarity of the texts under M. (This is of course a strong and unrealistic assumption; we will discuss how to deal with this fact further below.)

Given such an M, we can solve authorship attribution in a relatively straightforward manner.

For author classification, we would compare the questioned text Q to each of the known texts K_1 through K_n, and choose the author whose known texts are most similar to Q. If there is a near-tie, then we might have evidence of co-authorship. And if none of the known documents are sufficiently similar, and we have a large number and variety of known documents, we may conclude with some degree of certainty that Q's author is not one of the candidates.

This intuitive algorithmic schema is not, however, quite sufficient in practice. First, how do we devise a stylometric similarity measure that will have the desired correlation with authorship? Next, even given such a measure, what do we mean by "sufficiently similar"? How similar is similar enough? Third, how reliable can such a similarity measure be anyway? How can we know how reliable it is? Perhaps more importantly, since no similarity measure will be perfectly reliable, can we devise methods that are robust to not-perfectly-reliable similarity measures? How can characteristics of the known and questioned documents, such as number and length of documents and their genres, affect results? Finally, this overall framework does not tell us how to directly address the verification problem (we have no alternative candidates) or the profiling problem. We will now turn to outlining different specific algorithmic approaches which deal with these questions in a variety of ways.

Classifier learning

Perhaps the most straightforward approach is *classifier learning*, in which the set of known documents, each labeled with its correct author, is used as input to a classifier learning algorithm, whose output is a *classification model* m which outputs a predicted author for any input text it is given. A great variety of different classifer learning algorithms have been developed, many of which can be meaningfully applied to authorship attribution. Each has its own strengths and weaknesses, and the reliability of any particular method needs to be established for the particular task at hand. While a method should be chosen for its *plausibility* on the given problem based on the research literature, its reliability must also be evaluated on the available data, since the specifics of the scenario (the number of candidates and texts per candidate, the lengths, genres, diversity, etc. of the texts, and so forth) will affect accuracy, sometimes significantly (see Section 'Establishing Reliability' below).

The specific choice of classification algorithm, however, is less important than the composition of the corpus of known documents relative to the questioned document, and the choice of linguistic features by which to represent the character of a text. I briefly give an overview of classification learning here; for more detail about machine learning and how to use it, see (Domingos, 2012).

A classifier learning system C (see Figure 2) takes as input a set of known documents, each with a label (collectively the *training set*), represented as a set of document/label pairs, $\{\langle d_i, L_i\rangle\}$ – in authorship attribution, each label L_i is the known author of the corresponding document d_i. The output of C is a classification model m, which itself takes as input a document d and outputs a predicted label L. The goal is that m should classify new documents (not in the training set) with high accuracy. A key question therefore, which we will discuss in Section 'Establishing Reliability' below, is how to

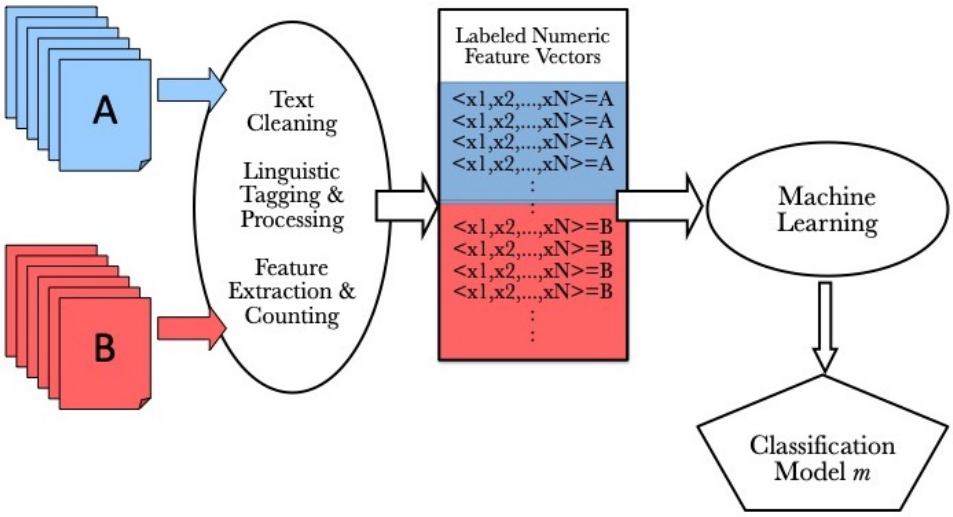

Figure 2. Flowchart showing the classification process (for two candidate authors). See text for details.

effectively evaluate C's ability to produce a high-accuracy m. Such evaluation is done using a separate *test set* of document/label pairs, where the labels assigned by m are to be compared to the 'correct' labels given in the test set.

Pitfall 0(b) (Evaluate methods on case documents) *Do not assume that a particular classification algorithm will work well for a given case, just because it has been shown to work in published research. If the texts used in that research differ qualitatively or quantitatively from those in the case, or the features used differ, results can be noticeably different. Whenever possible, you should evaluate the chosen method on the given documents in the case as well.*

Keep in mind that a classification model m will always give *some* answer for any text, so it is helpful (if possible) to use a method that can also give a (validated) measure of m's confidence in its answer. Such a measure, if reliable, can give a clearer picture of the strength of evidence provided.

Authorship Verification

In authorship verification, we seek to determine if a particular individual, A, wrote the questioned document Q. We are provided some known documents by A, but have no other candidates for the authorship of Q—the other candidates are everyone in the world other than A. A naive approach to verification would select some number of plausible alternative authors and show that A is more likely than any of them, using a classification approach. This will neither be reliable, nor convincing, since it is always possible that even if A's documents are closer to Q than any of these alternates, an even closer candidate B may be lurking just around the corner. (This is not an issue per se when a candidate author set is constructed based on the facts of the case.) So more sophisticated methods are needed.

Pitfall 0(c) (Verification $\neq$ 'more likely than known alternates') *If the question in the case is verification—whether or not a specific individual authored Q—it is not enough to just show that Q is a more likely author than an arbitrary set of alternate possibilities,*

since there is no guarantee that it is sufficiently broad to characterize the near-infinite alternatives.

Verification by classification

One scenario in which verification might be approached using classification techniques is when there is a known closed candidate set, but just one of the candidates is of interest. That is, the question is whether or not A wrote the document, and it is known that the author was either A or one of a small set of other candidates $B_1, ..., B_n$. Given known documents for all of these candidates, a straightforward approach would be to build a two-way classifier, deciding between similarity to A's known documents and the collected known documents of $B_1, ..., B_n$. If Q looks more like the former than the latter, than there is evidence to verify A. However, the fact that A's known documents will be less stylistically varied than those of $B_1, ..., B_n$ taken together can bias the process. This can be evaluated by also running n other similar tests, verifying authorship of B_1 versus $A, B_2, ..., B_n$, of B_2 versus $A, B_1, B_3, ..., B_n$, and so on. If the results are consistent, i.e., only one of the candidates is verified, the method can be considered potentially reliable in this case. But if many of them appear verified, the method has been shown to be unreliable in the given case.

Pitfall 0(d) (Test author vs. group classification for all candidates) *Even given a closed candidate set $A, B_1, ..., B_n$, verifying A's authorship by classifying A versus the other candidates is not a prima facie reliable procedure. You also need to probe the reliability of such binary classification by similarly verifying each of B_1, B_2, and so forth; unless all results are consistent, the original result cannot be considered reliable.*

Unmasking

One important type of scenario for which verification is the appropriate paradigm is when the potential author A is suspected of attempting to disguise their authorship. If A is at all competent at doing so, simple classification will likely fail, since they will include features that are highly uncharacteristic of their own writing, which will tend to confuse classification. This can also happen without deception, in some cases where known and questioned documents differ in extraneous ways such as genre or time of composition, that can introduce irrelevant but distinguishing features.

A method that has been shown to work well for such cases, despite this difficulty, is *unmasking* (Koppel *et al.*, 2007; Kestemont *et al.*, 2012). Suppose we have two sets S_1 and S_2 of documents (or sections of documents), where we know that each set has a single author, and we want to know if S_1 and S_2 have the same author. If there is no deception, then we could try to learn a classifier to distinguish S_1 documents from S_2 documents; if an accurate classifier can be learned, then the author is likely different, but if a learner cannot learn an accurate classifier, the sets are stylometrically indistinguishable, and so are likely by the same author. Obviously, this method will not work in the case of deception, since the (lying) author will have added artificial features to distinguish the document's style from their own, and the classifier will use them and get high accuracy.

The *unmasking* method unmasks these features by learning a sequence of classification models. After learning the first, cross-validation accuracy is measured (see below), and the features that contributed most to determining the classifications are removed from consideration. (Deception-based features are likely to be such strong features by

their nature.) Then learning is repeated with a reduced feature set, and accuracy measured. Again, strong features are removed, and learning with accuracy measurement repeated. This process is repeated a number of times, giving a sequence of generally declining accuracy values (an *unmasking sequence*), as more and more features are removed. However, if the case is one of deception, and the two document sets have the same author, we expect accuracy to dip sharply after a small number of iterations, once nearly all the deceptive features have been removed. This will not occur if the sets of documents do have different authors, rather accuracy will slowly decline over the entire range. By comparing the unmasking sequence of interest to others known to be for different authors, the existence of a significant dip can be verified directly.

The impostors method

Another method that addresses author verification is the *Impostors Method* (Koppel *et al.*, 2012a; Seidman, 2013; Stover *et al.*, 2016; Potha and Stamatatos, 2017). A key advantage of this method is that it does not rely on cross-validation like unmasking, and so requires much less data to work. The impostors method takes the questioned document Q and a known document K authored by the suspect author A, and determines the strength of evidence that Q and K share an author. The procedure works by analogy to a police lineup: In addition to Q and K, a set of *impostors* I_i is put together comprising documents by authors other than A which are as similar as possible in other ways to K and Q. The idea is that if the similarity between Q and K is more than that between Q and the impostors, then it is likely that Q and K share authorship. The impostors thus serve to normalize the similarity measure, telling us how similar we expect random pairs of documents to appear. The greater the number of independent impostors, the stronger the evidence is.

Pitfall 0(e) (Use enough impostors, similar to Q and K) *Use a sufficient number of impostors, and use impostors that are as similar as possible to both Q and K in all ways other than authorship.*

It is still possible, however, that Q and K are most similar by coincidence. Hence the full impostors method runs a large number (usually 100) trials, in each of which only a random subset of features is used for computing similarity. This way if the similarity of Q and K is only a coincidence, it will not often recur. So if Q and K are more similar than Q and any impostor in a large number k of these trials ($k > n$ for some threshold n), the evidence of coauthorship can be considered to be reliable. (Note that if $k < n$ that is not evidence against coauthorship, just the failure to make a positive attribution.) The choice of n will determine the false-positive and false-negative rates of the method—higher n will mean fewer erroneous attributions, but more missed attributions, and a lower n the reverse.

In published tests under laboratory conditions, and attribution threshold of $n = 20$ (out of 100 trials) gives false positive rates of below 10% (Koppel and Winter, 2014), but as for all methods, it is always advisable to test the impostor method on the texts of the case at hand. This can be done given sets of known documents by the suspect and other similar authors, considering how many same-author pairs are properly attributed and how many missed, and how many different-author pairs are improperly attributed and how many are not. This will give estimates of the false-positive and false-negative rates, which can be calibrated for how conservative a result is desired. For evidence, a conservative result, which has a low false-positive rate, is desirable, so that if an attribution is

made, it can be considered reliable. For investigations, a higher false-positive rate may be acceptable, if it lowers the likelihood that the actual author will slip through the net.

Pitfall 0(f) (Consider false-positive/false-negative tradeoff) *Consider whether your case requires conservative (only-if-very-sure) attribution, and thus a higher threshold for attribution.*

Pitfall 0(g) (Determine thresholds before testing) *Determine the threshold based on the literature and based on calibration tests on known documents* before *score attributions for Q, to avoid choosing a threshold that fits the results rather than interpreting the results based on a threshold.*

Finally, since the impostors method relies on statistics of large numbers, texts must be relatively long; the overall feature set must also be large to support many trials with random subsets.

Pitfall 0(h) (Use long documents and many features for the impostors method) *A rule of thumb based on research investigations is that texts should be around 2000 words long or longer to ensure reliability. (Shorter texts can be used if necessary, but reliability will degrade as shorter texts are used.) As well, since the impostors method performs many trials with different random subsets of an overall feature set, the full feature set must be relatively large (over 1000 features in general) to ensure sufficient variability among the subsets.*

Visual attribution

Since, as noted above, all computational attribution methods rely in some fashion on measuring some kind of similarity between documents, we might think of dispensing with fancy algorithmic attribution methods such as those we just discussed, and instead producing a visual representation of the stylometric relationships between documents and then visually determining which candidate author Q is most similar to, if any. This would be straightforward if, say, there were only two relevant linguistic features, so that every document would be represented as a pair of numbers (x, y) corresponding to the relative frequencies of the two features. Then we could plot all known documents on a graph, as in Figure 3, and determine the location of Q (also as a pair of numbers) compared to the known documents for each candidate. If Q's point is clearly in the 'cloud' of points for a particular candidate (as is the black circle in the figure), that gives good evidence for an attribution, and if it is far from any candidate documents' points (as is the black triangle in the figure), no attribution can be made (and possibly we have prima facie evidence to deny any candidate authorship if we can show that the known documents cover the full range of all the candidates' writings).

The trick, of course, is that textual style cannot be adequately represented by two numbers, however computed. Any plausible set of stylometric features will have dozens, if not hundreds, and so if we are to plot the known and questioned documents in two dimensions, we need to somehow boil a large number of dimensions down to two. Fortunately, there are standard statistical methods for doing so, which have been applied to authorship attribution.

The oldest and most standard such technique is *principal component analysis* (PCA). This rotates a set of numeric vectors in a multidimensional space to find axes such that the data distribution along each axis is statistically (linearly) uncorrelated with those

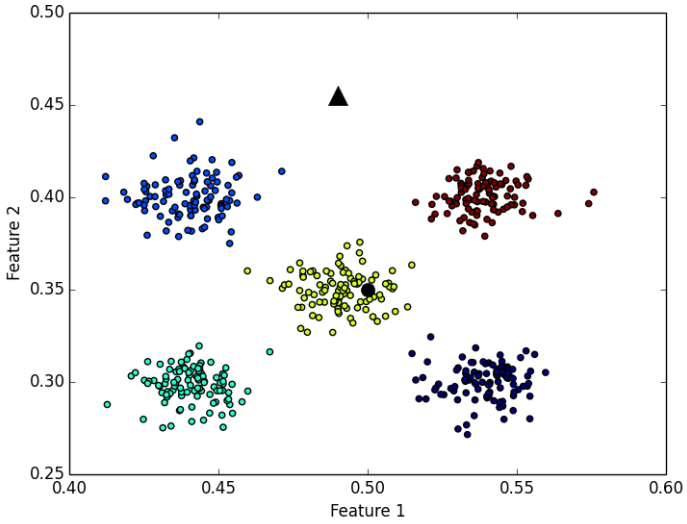

Figure 3. Simulated two-dimensional visualization of known documents from five clearly separated candidate authors shown in different colors, with two hypothetical questioned documents shown as a black circle and triangle.

along other axes (Abdi and Williams, 2010). These axes, called principal components, are ordered in descending order of how much data variability each contains. Hence, the first two components will give the data the widest spread of any two dimensions we could choose, and thus provide arguably the best two-dimensional representation of the data. For example, in his analysis of the authorship of the 15th Book of Oz, José Binongo (2003) plots known segments of Oz books known to be authored by the two candidate authors, L. Frank Baum and Ruth Plumly Thompson, per their first two principal components; we reproduce his figure in Figure 4. In this case, the known documents can be separated between the candidates perfectly using just the first principal component; we note that this level of clarity is very rare in practice.

Another technique for plotting high-dimension data in two dimensions is *multidimensional scaling* (MDS), which seeks to find an embedding of data points in two dimensions which maintains the relative distances between points, as much as possible (some distortion is inevitable, of course). MDS has been used similarly to PCA in authorship attribution research (López-Escobedo *et al.*, 2016). The techniques will give somewhat different results, as they are based on different definitions of what constitutes a 'good' reduction of the data to two dimensions, but the considerations and caveats for properly using them are similar.

The key such consideration is the deceptive simplicity of a scatter plot such as that in Figure 4. Visual inspection gives a clear answer—if Q falls on one side it was written by A, and if on the other side, B. The figure hides the complexity and statistical assumptions behind the result. The same procedure carried out on a slightly different set of documents, or on the same documents with different features, can give significantly different results. Using a different set of relevant documents also might. These possibilities need to be considered and ruled out, instead of simply relying on the force of visual clarity that the figure provides.

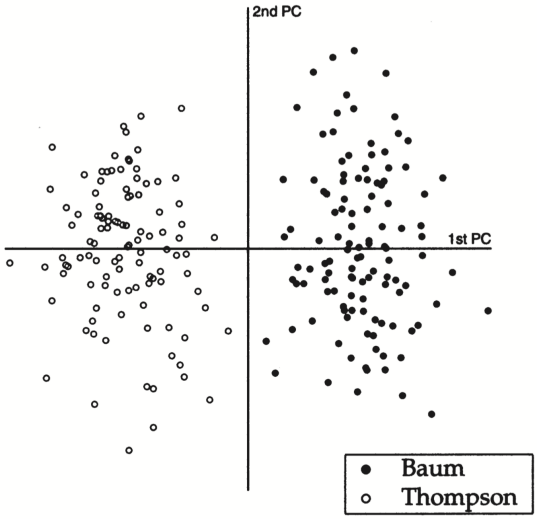

Figure 5. Baum vs. Thompson.

Figure 4. Texts by Baum vs. Thompson, plotted by first two principal components (Figure 5 of Binongo, 2003).

Pitfall 0(i) (Sensitivity testing for visualizations) *Dimensionality-reduced visualizations rest on complex assumptions and algorithms—don't just rely on visual clarity. You should probe how sensitive plots are to changes of features, similarity measures, and document sets before relying on them.*

That said, if a reliable two-dimensional plot can be constructed that gives a meaningful and useful answer, it can be very useful in making analysis results comprehensible to the judge and jury.

Clustering

One of the main goals of these visualization-based techniques is show how (or whether) the known documents divide into clear clusters by authorship, so that Q's authorship can be attributed by ascertaining which cluster it best belongs to. This idea can be implemented directly by using one of a number of *clustering algorithms* (Han *et al.*, 2011: Ch. 10) and see Berry and Castellanos (2004) and Xu and Tian (2015), which automatically divides known documents into a set of clusters, according to some criterion for the quality of such a division. Clustering is an *unsupervised* method of analysis, which does not use information about the authorship of the known documents to divide them into sets of stylistically similar documents. The idea is that if such sets correspond to specific authors, then the clustering has captured the stylistic correlates of authorship in the corpus, and the cluster identity of the questioned document is a likely indicator for its authorship.

Clustering has a long history of use in authorship analysis, as in Holmes and Forsyth's pioneering study of the Federalist Papers (1995) and Burrows's later application of the method to literary analysis of poetry and prose (2004). Cluster analysis for forensic authorship analysis may be less reliable, though, due to shorter text lengths and smaller corpora; the reliability of cluster analysis for literary texts has also been questioned (Hoover, 2003).

Even when considered effective, the results of clustering are highly sensitive to experimental parameters, such as the number and types of features, the distance measure used to compare feature vectors, and the way distances are aggregated to compare clusters with each other (Jain *et al.*, 1999; Halkidi *et al.*, 2001; Zaïane *et al.*, 2002). This difficulty can be somewhat mitigated through recent techniques that build consensus clusterings, combining information derived from many different parameter settings (Eder, 2017), but without a sensitivity analysis results cannot be considered reliable, just as noted above for visualizations.

Pitfall 0(j) (Sensitivity testing for clustering) *Clustering results can vary greatly depending on system parameters. Probe how sensitive results are to changes of features and other parameters before relying on them.*

Establishing Reliability

To evaluate an attribution method's reliability, we need to run it using some known documents for training and then test the result on new data for which we also know the correct answers (a *test set*). Note that the testing data must comprise different texts than the training, since it would be trivial (and meaningless) to get perfect accuracy on the training, simply by memorizing it.

Pitfall 0(k) (Ensure disjoint train and test sets) *If you test a model on the same documents used for training it, estimated accuracy will be considerably higher than you can expect for the questioned document. Make sure that training and testing are done on* different *documents.*

A difficulty, of course, is that to get an accurate model, we need as much training data as possible, but the available labeled data is usually limited. In experimental research, we gather as large a set of documents with known authors as possible, so that some can be used for training and some for testing, while in typical operational scenarios, the number of known documents is more limited. Regardless, only occasionally, even in research, do we have a truly enormous number of texts, and so we need to use those we have efficiently. The standard method to do this is *cross-validation* (Alpaydin, 2009), in which the available labeled data is divided randomly into a number (k) of equal-sized subsets $S_1, ..., S_k$, called *folds*, and k train/test evaluations are carried out. First, we apply a learning method C to build a model, training on $S_1, ..., S_{k-1}$ and test its accuracy on the last fold S_k. Then, we train on $S_1, ..., S_{k-2}, S_k$ and testing on the remaining fold S_{k-1}, and so forth, repeating the process a total of k times. The average of the k accuracy figures is then used as an estimate of the expected accuracy of C's learned model for future data. Note that cross-validation thus is able to use all of our labeled data for testing, while ensuring that at no time does it test a learned model on any of the data that was used for training it.

Even with cross-validation, however, you may have very few known documents in a given case, perhaps only two or three (or even just one) from each candidate author. In such a case, if the documents are long, one might consider increasing the number of training texts by splitting each document into sections. (See below for a discussion of text length.) Since the style within a particular document may vary slightly between sections of the document, this strategy can lead to more accurate models being constructed. However, cross-validation needs to be modified so that a model trained on part of a document is never tested on other parts of the same document. If it were, we could

not know if accuracy was due to detecting the authorship of the test text or due to the simple fact that they are from one document—about the exact same topic, in the exact same register and genre, for the exact same audience, etc. Hence, in this case, the split of the known texts must be done such that all sections of a single document are in the same fold, to avoid this problem.

Pitfall 0(l) (Don't train and test on sections of the same document) *If known documents are split into multiple sections, increasing the number of training texts, a model trained on some sections of one document cannot be tested on other sections from the same document. Thus all sections from a given document must be in the same fold when doing cross-validation.*

Another possible way to overcome the paucity of data would be to use other documents with known authors, other than the known documents in the case, to evaluate the method or to supplement those documents. The danger here is that if these documents are stylistically different from the documents in the case, whether in terms of register, genre, sociolect, discourse community, etc., the comparison may be invalid. Using results on other datasets can be used to argue for the plausibility of the method for the given case, but attention must be paid to the question of how similar the kinds of documents are to each other, and appropriate caveats attached. Best in such a case is to be able to point to multiple such tests that give consistent results. However, simply adding a number of unrelated documents to known documents from the case, to construct a larger training set, is likely to lead to results that cannot be trusted.

Pitfall 0(m) (Keep training set internally consistent) *Attribution accuracy can depend on the other influences on document style for training and test texts, and thus:*

- *Evaluations on documents not from the current case must be considered relative to the similarities and differences of the provenance of those documents to those available in the case, and*
- *External documents should not be mixed together with case documents to make a larger training set. The differences will lead to unreliable evaluation results.*

A subtle, and surprisingly important, question is raised when feature selection is done. In feature selection, a very large number of potential features, such as wordforms or part-of-speech n-grams, is whittled down to a manageable size by computing some measure of each feature's usefulness for classification and keeping the 'best' k features, or all those that pass a threshold. Such measures evaluate how well an individual feature can distinguish authors from each other; a variety of statistical measures exist such as information gain (Quinlan, 2014), chi-squared statistics (Moh'd A Mesleh, 2007), etc. Any such measure must be computed over labeled training data, wherein lies the danger. If features are selected based on an entire labeled corpus, and then learning (on the reduced feature set) evaluated through cross-validation, the test documents have actually been used in the training process, since feature selection is part of training. This is a very common error that is easy to fall into, but one which can lead to surprisingly misleading results. If this is done, evaluation results often greatly overestimate the accuracy of the classification method, which may appear accurate but turn out to be useless on new data.

Pitfall 0(n) (Don't use test data in feature selection) *If you use feature selection, make sure that selection measures are computing only over training data during evaluation, and not on test data. Otherwise you will overestimate the accuracy of your method.*

It is important also to note that accuracy itself is not a simple and unproblematic notion. If many more documents are available for one candidate author X than for others, high accuracy might be obtained simply by predicting that *all* documents were written by X. For example, if 70% of the known documents are by X and only 30% by other authors, this would give 70% accuracy. However, the method is clearly useless and meaningless, though it seems somewhat accurate based on the numbers. A more fine-grained evaluation is obtained by using two different "accuracy" measures—*precision* and *recall*. Precision measures, for each candidate author A, what fraction of the documents that the model m predicts are written by A were actually written by A. Recall, on the other hand, measures what fraction of the documents actually written by A were predicted by m to have been written by A. In our example of a dumb attribution method above, the precision for author X would be 70%, but for other authors would be undefined (since no predictions are made for them); recall for X would be 100%, but for other authors would be 0% (since they are never predicted). Thus we see how by looking at both precision and recall we get a better picture of the actual performance of the method.

Pitfall 0(o) (Use precision and recall for evaluation) *Simple accuracy as a measure can be affected significantly by imbalance in numbers of known documents for different candidates, and unreliable methods may appear reliable. Better is to calculate both precision and recall for each candidate author. This will show if all authors are treated equally by the learned model or if results are biased in one way or another.*

The harmonic mean of precision and recall, called the "F1 measure," is often used to give a single numeric metric for performance of text classification or information retrieval systems. It is better to use both precision and recall, however, for a couple of reasons. First, depending on the scenario, either precision or recall may be more important—averaging them loses clarity as to the import of the results. Second, in many realistic situations, high F1 can be obtained by methods that provide no useful information (Lipton *et al.*, 2014).

Corpus Composition

In all applications of authorship attribution, we must start with a *questioned document* (or set of documents) Q, whose authorship is to be determined, and a set of *known documents* K, which are reliably known to have been authored by the candidate author(s). For some attribution methods, particularly when dealing with open-set attribution or verification, a set of *comparison documents* C is also used, comprising documents by non-candidate authors as impostors or to provide background calibration for determining what level of stylometric similarity indicates coauthorship in the case.

When inferring authorship based on stylometric comparison of different texts it is essential to keep in mind the multiplicity of factors that can influence the stylistic character of a text (see Figure 5). There are no stylometric features that uniquely indicate author identity, hence care must be taken to rule out alternative explanations for stylometric similarity between two texts. As an extreme example, suppose Q is a corporate contract, and the question is which of two authors, A1 and A2, drafted it. If we are given one known document from each, K1 and K2, respectively, where K1 is a contract, and K2 is a personal email, the fact that Q is more similar to K1 than to K2 says nothing about its likely authorship, as the similarity is easily explained by register and genre.

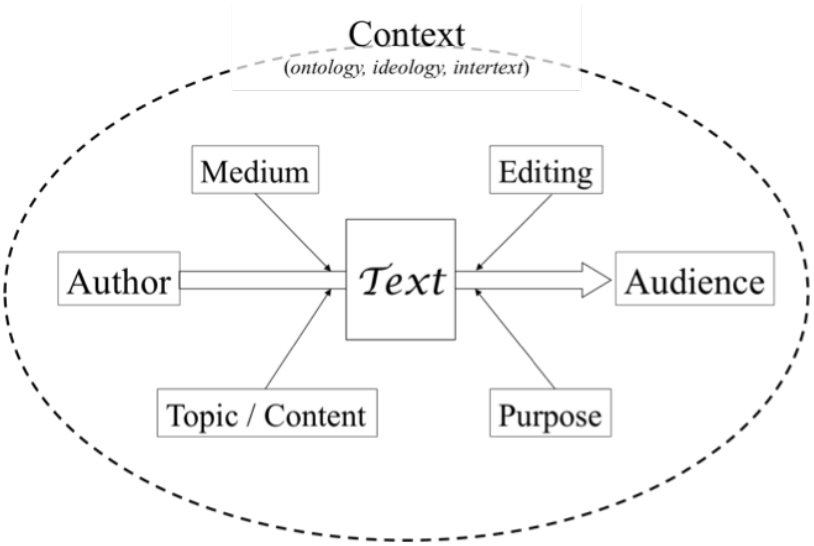

Figure 5. Summary of factors contributing to the precise form of a text (after Figure 5.1 of Argamon and Koppel, 2010). The Author seeks to express some Content about a Topic in a text via some Medium for some Purpose directed at some intended Audience. There may be Editing that affects the style and content of the text. The larger context within which the text's production is embedded also affects what text is produced, the relevant *ontology* assumed, the *ideology* encoding potential and actual social roles of the Author and Audience, and the *intertextual* relationships of the new Text with other texts that came before.

Ideally, all the documents, Q, K, and C, should be as similar as possible in all ways other than in authorship; this is the best way to ensure that inference to authorship cannot be explained by other factors. However, such a level of experimental control, exercised in laboratory research, is rarely if ever possible in the forensic context. Known documents are limited to whatever documents can be obtained for the candidate authors—there may be very few, and those that are available may be from different genres and registers from Q and from each other. It is critical to keep in mind that there are **no known** stylometric features that vary with authorship and do not vary with genre, register, topic, or other style-influencing factors (collectively, if vaguely, *text type*). Thus any differences in text type within the corpus must be accounted for, either by experimental control (which as noted is difficult to achieve in forensic cases), or by analytic procedure (see Section 'Algorithmic Attribution for discussion of how some methods can deal with differences in text types).

Pitfall 0(p) (Control corpora for text type) *If possible, ensure that all documents to be compared are of the same, or very similar, text types (genre, register, topic). If this cannot be assured, be very clear about the similarities and differences in text type and their likely influence on stylometric comparisons.*

Pitfall 0(q) (Exercise caution when Q differs in text type from K) *When Q differs in text type from known documents in K, and when known documents by different candidate authors differ in text type, consider carefully to what extent similarity judgments might be attributed to text type, as opposed to authorship, and could be misleading.*

In addition to controlling for text type, any method that relies on statistical analysis of textual features, as do the computational attribution methods discussed in this article,

must also control for text length. One is tempted to assume that the relative frequency of a given feature will be roughly the same no matter the length of the text. However, this is not the case. Common features will tend to drop in frequency as a text gets longer, due to the introduction of new vocabulary (cf. Zipf's law (1935)). See, for references, the frequencies of the words 'the' and 'you' in texts of different lengths in Figure 6—after an early rise, frequencies tend to drop until the text is long enough to give a near-constant frequency. Since forensic texts tend to be short, this variability is important to account for. Hence comparison of texts should be of segments of approximately equal length—if Q is 600 words long, comparing it to K_1 of length 600 words and K_2 of length 2500 words will not be a fair comparison, as we expect K_2 to have noticeably different frequency statistics from Q on general principle having nothing to do with authorship. Better is to use excerpts of (near-)equal length from all the documents to be compared.

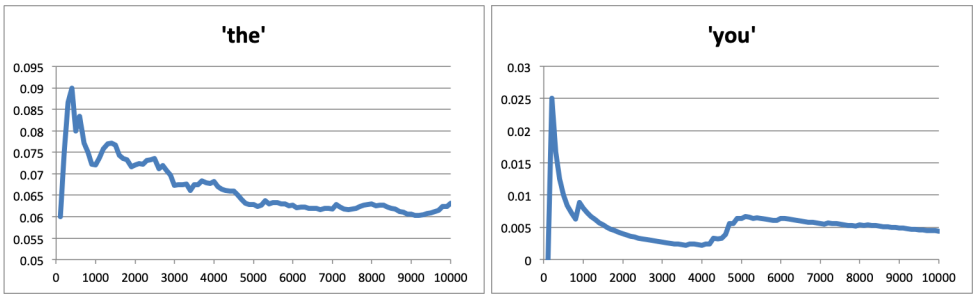

Figure 6. Frequencies of the words 'the' and 'you' in prefixes of different word lengths from the concatenated Congressional Record of the 104th–109th Congresses (Gentzkow and Shapiro, 2013).

Pitfall 0(r) (Control comparisons for text length) *Whenever possible, ensure that texts being compared are of nearly the same length, since estimates of lexical frequencies vary based on the length of the text (due to Zipf's Law). Such control is less critical when using syntactic features.*

Once we consider segmenting documents, however, we must contend with discourse structure—different sections of many kinds of document have different characters. This is true for many genres of text, including correspondence, suicide notes, contracts, essays, and more. Hence segments should respect boundaries between recognizable sections of each document, instead of comprising arbitrary segments of text of a given length. Also, assuming comparison is being done between texts of similar genre (recommended whenever possible), comparison should be between similar sections. Hence, the initial segment of one (say) letter should be compared against the initial segment of another, not its final segment. Inconsistency in this regard can lead to similarity judgments that are misleading.

Pitfall 0(s) (Control comparisons for discourse structure) *To the extent possible, ensure that text segments to be compared come from comparable portions of their respective documents. Ideally this would be based on a genre-relevant decomposition of the documents, but this can usually be approximated by using portions from the same relative positions in the respective documents (beginning, middle, end).*

The discussion above assumes that every document has a single author. While in some cases (e.g., ransom notes) this is a reasonable assumption, it is not always realistic. Editorial influence in published work can influence textual style, and some genres of text, such as contracts, are inherently multiply authored, due to collaboration and text reuse.

Pitfall 0(t) (Consider multiple authorship) *Consider directly the possibility that Q or known documents have multiple authorship or strong editorial influence. If single authorship can be assumed, make the assumption explicit and justify it.*

If single authorship cannot be strongly assumed, ideally plausibly multiply authored documents should be removed from consideration; however, this is rarely possible. Hence analysis must take the possibilities of collaborative authorship and editorial influence. How this is done depends greatly on the specifics of the case, but some general principles can be sketched.

If coauthorship of the questioned document Q is suspected, one approach is to attempt attribution separately for different sections of the document, which should detect if different individuals were primarily responsible for different sections. Except in cases where natural divisions are available (e.g., for plays, which divide into acts and scenes), overlapping sections should be chosen since we do not know in advance which portions may have been written by different people. The same idea can be applied to allowing for co-authorship of known documents, for attribution methods that treat known documents separately, so that each document section is compared to Q in isolation. In this case, it may be that Q (or a section thereof) matches only some sections of a known document, giving reason to believe that the known document may have multiple authors, and that Q may be attributable to whoever wrote those sections.

The likelihood of multiple authorship can also be directly tested by comparing the style of a document's sections to each other (Glover and Hirst, 1996; Graham *et al.*, 2005; Rybicki *et al.*, 2014; Stamatatos *et al.*, 2016). If different sections appear to show different authorial styles, they should be treated as separate units of analysis. If a known document seems to be multiply-authored, a conservative approach would simply remove it from consideration, provided that there are sufficient other known documents to proceed with the analysis.

Pitfall 0(u) (Segment documents to test and control for multiple authorship) *If multiple authorship cannot be ruled out, consider segmenting Q (and known documents) to be separately attributed. Stylistic comparison of segments of the same document can also be used to estimate the likelihood of the document being multiply authored.*

Segmenting documents will not, however, help us with the possibility of editorial influence, where authorial style is directly overlaid with other stylistic characteristics. In many cases, of course, the likelihood of editorial changes is virtually zero, as for ransom or suicide notes, but in cases involving published or institutional documents, this possibility is much more likely. Such influence may be from an editor's individual style, or from the imposition of a 'house style' on the document. Note that the implications for attribution of editorial influence are different when considering the questioned document or the known documents.

If Q's style may have been significantly affected by editorial changes, it will lessen the likelihood that any given candidate author is a strong match, since Q will bear a mixture of stylistic characteristics. Thus, if, nonetheless, just one candidate author is a strong match, the value of the evidence will be at least as large as had there been no editorial interference. However, it will be impossible to distinguish between attributed authorship and editorship—if candidate author A is a good match for Q, we cannot know if A was the author without external evidence that A was not the editor.

On the other hand, if editorial influence is suspected among the known documents, spurious similarities may be found with Q, invalidating analytic conclusions. One way to control for this, when multiple documents from each candidate author are available, is to compare known documents to each other. Each known document K is taken in turn as a questioned document, and attributed based on the remaining known documents. If editorial influence is minimal, we expect each known document to most likely be attributed correctly. If most are, but a small number are not, this may indicate editorial interference with those, and reason to exclude those known documents from consideration.

Pitfall 0(v) (Test for likelihood of editorial interference) *If significant editing of known documents cannot be ruled out, test for stylistic consistency among documents of each candidate author, and remove those that do not fit in with the rest.*

The above discussion assumes editorial influence varies for different documents. If the same editorial influence obtains for (say) all known documents by a single candidate, those documents may be stylistically consistent without clearly reflecting the style of the author—they may instead reflect the editor's style or a mixture of the two, without revealing an inconsistency. In such a case, we cannot reliably distinguish attribution to the candidate or to the editor.

Feature Extraction

We now consider the different sorts of textual features that are typically used in computational stylometric analyses, for authorship attribution as well as for others. Choice of such features must balance three considerations: their linguistic significance, their effectiveness at measuring true stylometric similarity, and the ease with which they can be identified computationally. Some potentially useful and linguistically meaningful features may not be easily (or at all) identified accurately by existing computational techniques. For example, metaphor use may be a useful feature for authorship analysis, but current automated metaphor identification methods are not accurate enough to rely upon.

Statistical complexity

The earliest work in stylometrics sought statistical measures invariant across documents by a single author but vary between authors. A great variety of such measures have been proposed, such as average word or sentence length (Fucks, 1952; Brinegar, 1963; Yule, 1939) and more complex statistics using type/token ratios and numbers of *hapax legomena* and the like, such as Yule's (1939) K, Sichel's (1975) S or Honore's (1979) R. However, no such measures have proven to be reliable for authorship attribution (Burrows, 1992; Grieve, 2007).

Pitfall 0(w) (Complexity measures are not reliable alone) *Overall measures of textual or linguistic complexity are not generally reliable for authorship attribution. Hence they should not be used except together with other features, if they increase a method's reliability. This must be demonstrated by empirical testing.*

Lexical choice

Lexical choice is a key dimension of variation between individual authors, who exhibit statistical preferences for different words that can be used in particular contexts. There are different kinds of feature sets built on this notion, as discussed below.

Function words

One of the oldest and most generally reliable feature sets used in stylometric authorship attribution is *function words*, used at least since Mosteller and Wallace's landmark study of the *Federalist Papers* (1964). Function word use (a) does not vary substantially with topic (but does with genre and register) and (b) constitutes a good proxy for a wide variety of syntactic and discourse-level phenomena. Furthermore, it is largely not under conscious control, and so should reduce the risk of being fooled by deception (Chung and Pennebaker, 2007).

Function word lists used in English are typically up to a few hundred words long and include pronouns, prepositions, auxiliary and modal verbs, conjunctions, and determiners, as well as numbers and interjections, even though they are not function words, since they tend to vary with authorship and are mostly topic-independent. The function words available for use in different languages will vary of course, and for synthetic languages will likely be incomplete and need to be supplemented by morphological analysis. Results of different studies using somewhat different lists of function words have been similar, indicating that the precise choice of function words is not crucial. Discriminators built from function word frequencies often perform at levels competitive with those constructed from more complex features.

Pitfall 0(x) (Use morphological analysis on synthetic languages) *Function word lists in synthetic languages will likely miss many important features of the idiolect, so morphological analysis is needed to extract a more complete set of features.*

When using function words for authorship attribution, attention must be paid to the fact that genre and register variation in the corpus will also affect function word frequencies. For example, pronouns (particularly first and second person) are much more frequent in narrative text than in informative text. Depending on the analysis methodology, some classes of function words may need to be removed from consideration.

Pitfall 0(y) (Filter function words based on genre and register) *Frequencies of many function words will vary greatly between different genres and registers of text, and so appropriate methods or controls need to be applied if the corpus must comprise diverse text types. This may involve removing some function words from consideration. All such controls must be validated empirically on the data.*

Content words

Other aspects of lexical choice variation are not captured by function word use. For example, one candidate author may prefer to use words like 'start' and 'large', where another may prefer 'begin' and 'big' (Mosteller and Wallace, 1964; Koppel *et al.*, 2006, 2009). This sort of pattern can be analyzed by modeling the relative frequencies of content words. Typically very rare words and those with near-uniform distribution over the corpus of interest can be omitted (Forman, 2003), so that a set of several to ten thousand words may be used. Content words, however, require even tighter experimental care and control, since their frequencies will vary with topic, as well as with text type. This may lead to both false attributions and to missing valid attributions, depending on how such irrelevant dimensions of variation may influence attribution.

Pitfall 0(z) (Using content words requires tighter corpus control) *Content words may indicate topic more strongly than authorship, so tests using them need tight controls*

on topic of corpus documents, or methods that can be shown to be stable in the face of topic differences. Examining the features that the analysis identifies as key to the attribution should be done to check if such interference is present.

Word embeddings

Using words as features for stylometric comparison, whether function words or content words, finds similarity by comparing occurrences of the exact same word. However, some words are more similar than other. Consider a comparison between the sentences "The President spoke about tariffs" and "The administration issued a statement about import taxes." The only words shared between them are "the" and "about," however, they are very similar. Significant semantic closeness is seen in the pairs (President, administration), (spoke, statement), and (tariffs, taxes), but is not taken into account by word-based methods. A popular way to generalize word comparison is to use a *word embedding*, which represents each word by a multidimensional numeric vector such that words that occur in similar contexts will have similar vectors. One of the most popular methods, Word2vec (Mikolov *et al.*, 2013), uses a neural network model to derive such embeddings, largely capturing semantic and syntactic connections between words such that similar words have nearby vectors. They show, for example, that vec[*king*]+(vec[*woman*]-vec[*man*]) $\approx$ vec[*queen*]. Recent development of *contextual* word embeddings (Devlin *et al.*, 2018; Peters *et al.*, 2018) give more precise word vectors for particular word occurrences, that are sensitive to context. These embeddings thus encode different word senses and parts-of-speech, giving a more fine-grained representation.

The hope of using such vectors for stylometric comparison, is to get more general and more precise measures of semantic similarity in lexical choice. Indeed, some recent research has shown word embeddings to give useful features for authorship analysis in research studies (Sari and Stevenson, 2016; Posadas-Durán *et al.*, 2017). Results seem fairly insensitive to what corpus was used to compute the embedding, provided it was large enough—standard embeddings trained on very large corpora are now easily available for such use. The main caveat when using word embeddings is that, just like content words, their occurrence is dependent on document topic, genre, and register, and so these factors need to be tightly controlled in any authorship analysis using them.

Pitfall 0() (Word embeddings encode topic dependence) *Word embeddings enable better determination of lexical similarity by generalizing beyond identity of word tokens. However, they share the properties of topic- and text type-dependence of content words, and analysis must be controlled accordingly.*

Syntax

Another category of style markers is the relative frequencies of different choices of syntactic structure, either measured directly, or by proxy via looking at occurrences of parts of speech. Different authors have different preferences for type and complexity of different constructs, and both absolute and relative frequencies of syntactic constructs have shown to be useful for authorship attribution (Baayen *et al.*, 1996; Stamatatos *et al.*, 2001; Gamon, 2004; Hirst and Feiguina, 2007). In all such cases, feature frequency is likely to be influenced by text type, and so experimental control is necessary (or text-type invariance needs to be demonstrated).

Pitfall 0() (Syntax also requires text type control) *Despite its facial and empirical topic independence, syntactic choice is* not *invariant to text type; different genres and registers have difference characteristic relative frequencies for various syntactic constructs. Hence full control for text type is necessary when using syntactic features as well.*

Extracting syntactic structure from text in English and most other European languages can be done accurately using current natural language processing tools, for texts in reasonably standard prestige dialect. These tools will have more difficulty on less formal text that includes orthographic and grammatical errors or variations, as well as on most languages outside the European mainstream.

Pitfall 0() (Understand accuracy of syntactic analysis tools) *Automated syntactic analysis tools vary in the accuracy of their output depending on the language (they are best for English and major European languages) and text type. They are particularly poor on informal texts. Their accuracy should be evaluated on texts of the same kind as the analysis corpus before use.*

A simple type of syntax-based feature is using relative frequencies of different parts-of-speech and of short part-of-speech sequences, e.g., "the fraction of common nouns that are immediately preceded by an adjective". A number of research studies have shown that such features can be useful in authorship attribution (Argamon *et al.*, 1998; Kukushkina *et al.*, 2001; Corney *et al.*, 2001; Koppel *et al.*, 2002; Koppel and Schler, 2003; Zhao *et al.*, 2006; Zheng *et al.*, 2006).

More complex automated parsing tools can be used to identify full syntactic structures, and compute the frequencies of noun and verb phrases or of relative clauses, for example. These have also been shown to work for authorship attribution in the research literature (Baayen *et al.*, 1996; Stamatatos *et al.*, 2001, 2000; Gamon, 2004; Halteren, 2007; Chaski, 2005; Uzuner *et al.*, 2005; Hirst and Feiguina, 2007).

Specific examples of such features are:

- N-grams of parts-of-speech: "determiner–adjective–adjective" or "common noun–common noun" (Argamon-Engelson *et al.*, 1998),
- Syntactic phrase categories: XYZ (Stamatatos *et al.*, 2001)
- Syntactic category bigrams: "coordinating conjunction followed by clause" or "name starting with proper noun" (Hirst and Feiguina, 2007), and
- Marked syntactic structures: "non-head-final noun phrase" (in English) (Chaski, 2005).

In most attribution studies, syntactic features are used together with lexical features, as syntactic features alone are not usually fine-grained enough to attain high accuracy.

Pitfall 0() (Evaluate if syntax is reliable for the specific case) *Consider well whether the syntactic features to be used are likely to be reliable for the kinds and numbers of texts in the corpus, and empirically test them. Use lexical features (e.g., function words) as well, if needed.*

Character n-grams

The relative frequencies of character n-grams (sequences of several characters), such as "ing", "auth", "opos", or the like, has been proposed as a feature set for attribution, subsuming lexical choice features (function and content words) and morphology (by

capturing many affixes). Such features have the big advantage of being largely language-independent (for non-ideographic writing systems); a number of research studies have shown their efficacy for attribution in various languages and contexts (Kjell, 1994; Clement and Sharp, 2003; Houvardas and Stamatatos, 2006; Ledger and Merriam, 1994; Grieve, 2007; Kešelj *et al.*, 2003; Peng *et al.*, 2004). Since they are sensitive to topic as well as text type, all of the concerns regarding function and content words apply as well to character n-grams.

Presenting Results

No text analytic method can conclusively prove who the author of a questioned text is—a good result is one which shows where the weight of the evidence lies, with respect to the authorship question at hand, and gives some measure of the strength of that evidence. An attribution result is one of two types: it may *rule-in* a particular candidate A as a likely author of Q, or it may *rule-out* a candidate B, tagging B as an unlikely author of Q. In both cases, one must be careful to determine and explain who the alternative authors are that A (or B) is being compared against (see the discussion in Section 'Task Formulation' above comparing open- and closed-set classification and verification tasks).

Pitfall 0() (No analysis can prove authorship) *Never claim that an analysis "demonstrates" authorship. The best that can be said is where the strength of the evidence points, compared to particular alternatives.*

Strength of evidence

If A is being ruled in as a likely author (or coauthor) of Q, the strength of the evidence will be that A's known documents K_A are particularly similar to Q, relative to known documents by other potential candidates and/or background authors representing the rest of the world. The metric for similarity needs to be calibrated, and that calibration shown, to show how similar is similar enough to determine likely authorship, and what the error rates, both false positive and false negative, are likely to be.

Pitfall 0() (Exhibit calibration on known documents) *Attribution measures for relevant documents with known authorship should be shown for calibration, to enable the jury to evaluate themselves the significance of your attribution results.*

When presenting quantitative results, particularly estimates of reliability of the analysis, it is important to do so in a way that avoids fallacies in interpretation. For example, suppose an analysis is performed to find the author of a questioned text Q from (say) a thousand candidates, and one candidate, X, matches Q such that the estimated probability of the match happening by chance is just one in a thousand. If that probability is presented as-is to a jury, their direct (and fallacious) conclusion may be that there is a 99.9% chance that X is the author of Q. This, however, is an instance of the prosecutor's fallacy (Thompson and Schumann, 1987). The actual probability of *some* candidate among the thousand reaching this level of match with Q by chance is $1-(1-\frac{1}{1000})^{1000} \approx 63.2\%$, thus, on its own, this is weak evidence for X's authorship indeed.

A less misleading presentation of evidential power of the attribution to X would be to present it in terms of Bayesian updating of the probability of the attribution given the new evidence (Berger, 2013), by giving the *Bayesian update factor* to the prior probability for X's authorship given the analysis:

$$\frac{P(\text{author=X}|\text{analysis})}{P(\text{author=X})}$$

This formulation directly shows how evidence adduced by the analysis should be combined with other available evidence to form a conclusion, and can be intuitively explained as an update to prior beliefs about the candidate.

Precise probability estimates are not always available, and such estimates often themselves rely on probabilistic assumptions. This can be most clearly expressed by giving a confidence interval, saying, for example, that the Bayes update factor is most likely between 1.5 and 6, so it is at least 50% more likely that X is the author given the analysis, and perhaps as much as six times more likely.

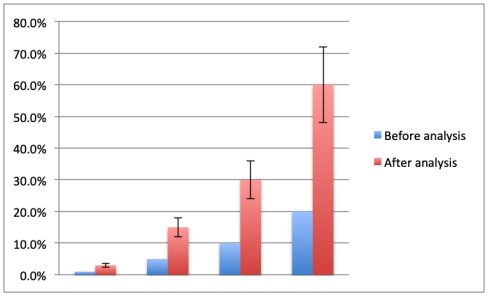

Figure 7. Example bar-graph showing belief updating for a Bayes update factor of $3 \pm 20\%$.

Visualizations can also be helpful. One way is to concretize these notions of probability—a bar graph (say) such as in Figure 7, showing how to update one's level of belief in authorship given the analysis. Another is to graphically show the similarity of Q to the known documents by different candidates (and comparison documents when relevant), as in Figures 3 and 4 above, for example. If this is done, care must be taken to address the potential pitfalls described in Section 'Visual attribution, and to explain how this was done, of course.

Opening the black box

In addition, whatever attribution method is used should not be treated as a black box that simply takes documents as input and outputs attributions (with confidence scores). The box needs to be opened up to show what features are doing the attribution work, that is, which features Q shares more with K_A than with documents not by A. This helps to establish the trustworthiness of the method, as well as give more detail to the evidential claim of authorship.

The same principle applies when ruling out an author B. In such a case, the claim is supported by the similarity of B's known documents K_B to Q being notably less than would be expected if B was an author of Q. Here, opening up the box means showing features that are shared between different texts authored by B, but that are not shared with Q.

Pitfall 0() (Show the features that support the analysis) *Do not treat an analysis method as a black box, but show what textual features it bases its result on. This is necessary to establish the strength and the basis of the evidence for authorship being adduced.*

Examining the features used by the algorithms to classify authorship is also essential as a check on the entire text-processing pipeline. It is surprisingly easy, when dealing with diverse input formats, for text preprocessing to let through tokens that are not part of

the actual text such as "page 3" or the name of an author in a page header, or the like. If such errors affect attribution, telltale features will show up, letting the analyst know to debug the text processing subsystem.

It should be noted that in the currently popular 'deep learning' techniques, as well as some others, it is not possible to directly determine what features are used to determine authorship. Indeed, explaining why a particular result was reached by such a model is, in general, an important unsolved research problem (Biran and Cotton, 2017; Samek *et al.*, 2017).

Concluding Thoughts

Computational authorship analysis methods can often allow reliable attribution even in cases where purely manual linguistic analysis is difficult or impossible, by statistical analysis of a very large number of subtle stylistic markers. However, establishing the reliability of a particular method for a particular case can be tricky, as it depends critically on many specifics of the case—one cannot simply rely on previous experience or experiments with the method. The list of potential pitfalls in this article should serve as guidelines for ensuring good methodology in developing computational authorship analyses, but the reader should always keep in mind that no such list can ever be complete. Expertise, experience, and careful human judgment must always be used and never supplanted by blind adherence to any predetermined methodology.

Summary List of Potential Pitfalls

References

Abdi, H. and Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4), 433–459.

Alpaydin, E. (2009). *Introduction to machine learning*. MIT press.

Argamon, S. and Koppel, M. (2010). The rest of the story: Finding meaning in stylistic variation. In *The Structure of Style*. Springer, Berlin, Heidelberg, 79–112.

Argamon, S., Koppel, M. and Avneri, G. (1998). Routing documents according to style. In *First International workshop on innovative information systems*, 85–92.

Argamon-Engelson, S., Koppel, M. and Avneri, G. (1998). Style-based text categorization: What newspaper am i reading. In *Proc. of the AAAI Workshop on Text Categorization*, 1–4.

Arun, R., Suresh, V. and Madhavan, C. V. (2009). Stopword graphs and authorship attribution in text corpora. In *2009 IEEE International Conference on Semantic Computing*, 192–196: IEEE.

Baayen, H., Van Halteren, H. and Tweedie, F. (1996). Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, 11(3), 121–132.

Berger, J. O. (2013). *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media.

Berry, M. W. and Castellanos, M. (2004). Survey of text mining. *Computing Reviews*, 45(9), 548.

Binongo, J. N. G. (2003). Who wrote the 15th book of oz? an application of multivariate analysis to authorship attribution. *Chance*, 16(2), 9–17.

Biran, O. and Cotton, C. (2017). Explanation and justification in machine learning: A survey. In *IJCAI-17 Workshop on Explainable AI (XAI)*, 8.

Brinegar, C. S. (1963). Mark twain and the quintus curtius snodgrass letters: A statistical test of authorship. *Journal of the American statistical Association*, 58(301), 85–96.

Burrows, J. (2004). Textual analysis.". *A companion to digital humanities*, 323–347.

Burrows, J. F. (1992). Not unless you ask nicely: The interpretative nexus between analysis and information. *Literary and Linguistic Computing*, 7(2), 91–109.

Cancedda, N., Gaussier, E., Goutte, C. and Renders, J.-M. (2003). Word-sequence kernels. *Journal of machine learning research*, 3(Feb), 1059–1082.

Chaski, C. E. (2005). Who's at the keyboard? authorship attribution in digital evidence investigations. *International journal of digital evidence*, 4(1), 1–13.

Chung, C. and Pennebaker, J. W. (2007). The psychological functions of function words. *Social communication*, 1, 343–359.

Clement, R. and Sharp, D. (2003). Ngram and bayesian classification of documents for topic and authorship. *Literary and linguistic computing*, 18(4), 423–447.

Corney, M. W., Anderson, A. M., Mohay, G. M. and de Vel, O. (2001). Identifying the authors of suspect email. *Communications of the ACM*.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55(10), 78–87.

Eder, M. (2017). Visualization in stylometry: Cluster analysis using networks. *Digital Scholarship in the Humanities*, 32(1), 50–64.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of machine learning research*, 3(Mar), 1289–1305.

Fucks, W. (1952). On mathematical analysis of style. *Biometrika*, 39(1/2), 122–129.

Gamon, M. (2004). Linguistic correlates of style: authorship classification with deep linguistic analysis features. In *Proceedings of the 20th international conference on Computational Linguistics*, 611: Association for Computational Linguistics.

Gentzkow, M. and Shapiro, J. (2013). *Congressional Record for 104th-109th Congresses: Text and Phrase Counts*. Rapport interne ICPSR33501-v2, University of Michigan, Ann Arbor, MI.

Glover, A. and Hirst, G. (1996). Detecting stylistic inconsistencies in collaborative writing. In *The new writing environment*. Springer, 147–168.

Graham, N., Hirst, G. and Marthi, B. (2005). Segmenting documents by stylistic character. *Natural Language Engineering*, 11(4), 397–415.

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

Halkidi, M., Batistakis, Y. and Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of intelligent information systems*, 17(2-3), 107–145.

Halteren, H. V. (2007). Author verification by linguistic profiling: An exploration of the parameter space. *ACM Transactions on Speech and Language Processing (TSLP)*, 4(1), 1.

Halvani, O., Winter, C. and Graner, L. (2018). Unary and binary classification approaches and their implications for authorship verification. *arXiv preprint arXiv:1901.00399*.

Han, J., Pei, J. and Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Hirst, G. and Feiguina, O. (2007). Bigrams of syntactic labels for authorship discrimination of short texts. *Literary and Linguistic Computing*, 22(4), 405–417.

Holmes, D. I. and Forsyth, R. S. (1995). The federalist revisited: New directions in authorship attribution. *Literary and Linguistic computing*, 10(2), 111–127.

Honore, T. (1979). 'imperial'rescripts ad 193–305: Authorship and authenticity. *The Journal of Roman Studies*, 69, 51–64.

Hoover, D. L. (2003). Multivariate analysis and the study of style variation. *Literary and Linguistic Computing*, 18(4), 341–360.

Houvardas, J. and Stamatatos, E. (2006). N-gram feature selection for authorship identification. In *International Conference on Artificial Intelligence: Methodology, Systems, and Applications*, 77–86: Springer.

Jain, A. K., Murty, M. N. and Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys (CSUR)*, 31(3), 264–323.

Kešelj, V., Peng, F., Cercone, N. and Thomas, C. (2003). N-gram-based author profiles for authorship attribution. In *Proceedings of the conference pacific association for computational linguistics, PACLING*, volume 3, 255–264.

Kestemont, M., Luyckx, K., Daelemans, W. and Crombez, T. (2012). Cross-genre authorship verification using unmasking. *English Studies*, 93(3), 340–356.

Kestemont, M., Stover, J., Koppel, M., Karsdorp, K. and Daelemans, W. (2016). Authorship verification with the ruzicka metric. In *Proceedings of Digital Humanities*, 246–249.

Kjell, B. (1994). Authorship determination using letter pair frequency features with neural network classifiers. *Literary and Linguistic Computing*, 9(2), 119–124.

Koppel, M., Argamon, S. and Shimoni, A. R. (2002). Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4), 401–412.

Koppel, M. and Schler, J. (2003). Exploiting stylistic idiosyncrasies for authorship attribution. In *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*, volume 69, 72–80.

Koppel, M., Schler, J. and Argamon, S. (2009). Computational methods in authorship attribution. *Journal of the American Society for information Science and Technology*, 60(1), 9–26.

Koppel, M., Schler, J. and Argamon, S. (2012a). Authorship attribution: What's easy and what's hard. *JL & Pol'y*, 21, 317.

Koppel, M., Schler, J., Argamon, S. and Messeri, E. (2006). Authorship attribution with thousands of candidate authors. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, 660: ACM.

Koppel, M., Schler, J., Argamon, S. and Winter, Y. (2012b). The "fundamental problem" of authorship attribution. *English Studies*, 93(3), 284–291.

Koppel, M., Schler, J. and Bonchek-Dokow, E. (2007). Measuring differentiability: Unmasking pseudonymous authors. *Journal of Machine Learning Research*, 8(Jun), 1261–1276.

Koppel, M. and Winter, Y. (2014). Determining if two documents are written by the same author. *Journal of the Association for Information Science and Technology*, 65(1), 178–187.

Kukushkina, O. V., Polikarpov, A. A. and Khmelev, D. V. (2001). Using literal and grammatical statistics for authorship attribution. *Problems of Information Transmission*, 37(2), 172–184.

Ledger, G. and Merriam, T. (1994). Shakespeare, fletcher, and the two noble kinsmen. *Literary and Linguistic Computing*, 9(3), 235–248.

Lipton, Z. C., Elkan, C. and Naryanaswamy, B. (2014). Optimal thresholding of classifiers to maximize f1 measure. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 225–239: Springer.

Lodhi, H., Shawe-Taylor, J., Cristianini, N. and Watkins, C. J. (2001). Text classification using string kernels. In *Advances in neural information processing systems*, 563–569.

López-Escobedo, F., Solorzano-Soto, J. and Sierra Martínez, G. (2016). Analysis of intertextual distances using multidimensional scaling in the context of authorship attribution. *Journal of Quantitative Linguistics*, 23(2), 154–176.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.

Moh'd A Mesleh, A. (2007). Chi square feature extraction based SVMs Arabic language text categorization system. *Journal of Computer Science*, 3(6), 430–435.

Mosteller, F. and Wallace, D. (1964). *Inference and disputed authorship: The Federalist.* Addison-Wesley.

Mostrous, A. (2013). JK Rowling unmasked as author of bestselling crime novel. *The Times (UK).*

Peng, F., Schuurmans, D. and Wang, S. (2004). Augmenting naive bayes classifiers with statistical language models. *Information Retrieval*, 7(3-4), 317–345.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, 2227–2237.

Posadas-Durán, J.-P., Gómez-Adorno, H., Sidorov, G., Batyrshin, I., Pinto, D. and Chanona-Hernández, L. (2017). Application of the distributed document representation in the authorship attribution task for small corpora. *Soft Computing*, 21(3), 627–639.

Potha, N. and Stamatatos, E. (2017). An improved impostors method for authorship verification. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, 138–144: Springer.

Quinlan, J. R. (2014). *C4. 5: Programs For Machine Learning*. Morgan Kaufman.

Rybicki, J., Hoover, D. and Kestemont, M. (2014). Collaborative authorship: Conrad, ford and rolling delta. *Literary and Linguistic Computing*, 29(3), 422–431.

Salton, G. and Lesk, M. E. (1968). Computer evaluation of indexing and text processing. *Journal of the ACM (JACM)*, 15(1), 8–36.

Samek, W., Wiegand, T. and Müller, K.-R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*.

Sari, Y. and Stevenson, M. (2016). Exploring word embeddings and character n-grams for author clustering. In *CLEF (Working Notes)*, 984–991.

Seidman, S. (2013). Authorship verification using the impostors method. In *CLEF 2013 Evaluation Labs and Workshop-Online Working Notes*: Citeseer.

Sichel, H. S. (1975). On a distribution law for word frequencies. *Journal of the American Statistical Association*, 70(351a), 542–547.

Stamatatos, E., Fakotakis, N. and Kokkinakis, G. (2000). Automatic text categorization in terms of genre and author. *Computational linguistics*, 26(4), 471–495.

Stamatatos, E., Fakotakis, N. and Kokkinakis, G. (2001). Computer-based authorship attribution without lexical measures. *Computers and the Humanities*, 35(2), 193–214.

Stamatatos, E., Tschnuggnall, M., Verhoeven, B., Daelemans, W., Specht, G., Stein, B. and Potthast, M. (2016). Clustering by authorship within and across documents. In *Working Notes Papers of the CLEF 2016 Evaluation Labs*, 691–715.

Stover, J. A., Winter, Y., Koppel, M. and Kestemont, M. (2016). Computational authorship verification method attributes a new work to a major 2nd century african author. *Journal of the Association for Information Science and Technology*, 67(1), 239–242.

Thompson, W. C. and Schumann, E. L. (1987). Interpretation of statistical evidence in criminal trials. *Law and Human Behavior*, 11(3), 167–187.

Uzuner, Ö., Katz, B. and Nahnsen, T. (2005). Using syntactic information to identify plagiarism. In *Proceedings of the second workshop on Building Educational Applications Using NLP*, 37–44: Association for Computational Linguistics.

Vilariño, D., Pinto, D., Gómez, H., León, S. and Castillo, E. (2013). Lexical-syntactic and graph-based features for authorship verification. In *PAN workshop at CLEF*.

Xing, Z., Pei, J. and Keogh, E. (2010). A brief survey on sequence classification. *ACM Sigkdd Explorations Newsletter*, 12(1), 40–48.

Xu, D. and Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science*, 2(2), 165–193.

Yule, G. U. (1939). On sentence-length as a statistical characteristic of style in prose: With application to two cases of disputed authorship. *Biometrika*, 30(3/4), 363–390.

Zaïane, O. R., Foss, A., Lee, C.-H. and Wang, W. (2002). On data clustering analysis: Scalability, constraints, and validation. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 28–39: Springer.

Zhao, Y., Zobel, J. and Vines, P. (2006). Using relative entropy for authorship attribution. In *Asia Information Retrieval Symposium*, 92–105: Springer.

Zheng, R., Li, J., Chen, H. and Huang, Z. (2006). A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American society for information science and technology*, 57(3), 378–393.

Zimmer, B. (2013). The science that uncovered J.K. Rowling's literary hocus-pocus. *The Wall Street Journal*.

Zipf, G. (1935). *The Psychology of Language*. Houghton-Mifflin.

Developing forensic authorship profiling

Andrea Nini

University of Manchester, UK

Abstract. *Current research into the task of determining the characteristics of an anonymous writer, authorship profiling, does not meet the demands of the forensic context, because of the lack of transparency of certain computational techniques, their requirements for large data sets, and, most importantly, since the strength of register variation does not guarantee that findings obtained in other registers will apply to forensic registers such as, for example, a threatening letter. The present article demonstrates how previously established findings related to stylistic variation in English for gender, age, and social class also apply to the kinds of texts often analysed by forensic linguists through an experiment involving 96 participants. These results constitute an example of linguistically-motivated profiling research and it is argued that the agenda to move from authorship profiling to forensic authorship profiling should be led by previously established knowledge of language variation.*

Keywords: *Authorship profiling, register variation, stylistics, threatening text, corpus linguistics.*

Resumo. *A atual investigação sobre a determinação das características de um escritor anónimo, a determinação do perfil do autor, não satisfaz as necessidades do contexto forense devido à falta de transparência de determinadas técnicas computacionais, dos seus requisitos para grandes "data sets" e, sobretudo, devido ao facto de a robustez da variação do registo não garantir que os resultados obtidos noutros registos sejam aplicáveis aos registos forenses como, por exemplo, uma carta de ameaça. Este artigo demonstra de que modo estudos prévios relacionados com a variação estilística em inglês relativamente ao género, idade e classe social também são aplicáveis aos tipos de texto muitas vezes analisados pelos linguistas forenses; para o efeito, realizou-se uma experiência que envolveu 96 participantes. Estes resultados constituem um exemplo de investigação na determinação de perfis linguisticamente motivada, defendendo-se que o plano de investigação para passar da determinação de perfis de autor para a determinação de perfis de autor forense deveria ser orientada por investigação prévia sobre variação linguística.*

Keywords: *Determinação de perfis de autor, variação de registo, estilística, texto de ameaça, linguística de corpus.*

Introduction

Authorship profiling is the task of determining the characteristics of an anonymous author, such as their demographic details, from the way they use language. Profiling questions can be of extreme importance at the investigative phases of, for example, a case involving an anonymous threatening letter or a ransom demand, when the list of suspects is too large. Despite this importance, the forensic linguistics literature on authorship profiling is very limited. Two ways of doing authorship profiling have emerged from forensic casework and research: (1) analysis of salient linguistic markers, and (2) analysis of writing style.

The first type of profiling is the application of sociolinguistic knowledge on a case by case basis to extract *ad hoc* linguistic features that are markers of a certain demographic background, as demonstrated in famous examples such as the *devil strip* case (Leonard, 2005), the Unabomber case (Shuy, 2014) or the *bad-minded* case (Grant, 2008). This type of analysis involves the linguist's experience in discovering dialectal or sociolinguistic features that can reveal clues about the background of the author.

In contrast, the second type of profiling consists in the analysis of the *stylistic variation* exhibited by the text as a whole. This analysis often involves the study of the frequency with which certain features are used, like the study of register variation (Biber, 1988) and takes as the unit of analysis the text itself. A *style*, as defined by Biber and Conrad (2009), is a variety of language associated with a particular author or social group as opposed to a situation which is constituted by linguistic features that are pervasive and frequent. It is therefore similar to a *register*, which is a variety of language associated with a particular situation, in terms of feature types that constitute it, but different in that styles are particular varieties of registers that characterise authors or social groups.

The current state of the art of authorship profiling reveals that research on the first type of analysis is virtually non-existent while the second type has become a sub-field of computer science and machine learning. It is indeed very difficult to systematise research for Type 1 profiling, as the type of markers that become important in a forensic case is often unpredictable. Analysis of Type 1 therefore relies almost entirely on the knowledge and intuition of the forensic linguist. Research on Type 2 profiling, on the other hand, has been developed by computer scientists applying machine learning techniques, for example, to automatically determine the gender or age of a writer (Argamon *et al.*, 2009). These systems usually work by taking as input an array of features, usually frequencies of words or characters, and using these arrays to train a machine to distinguish groups of texts that have been labelled already as, for instance, male or female. The outputs of these systems are the classification accuracies and, sometimes, the distinguishing features.

The fact that Type 2 authorship profiling is dominated by computer science can be quite problematic for forensic linguistics, since the needs of forensic linguists are often different from the needs of the users of computational applications. Computational authorship profiling is not necessarily interested in understanding the inner (linguistic) mechanisms of the machine, as long as the accuracy rates are outperforming previous models. This lack of linguistic understanding can however be problematic for a forensic linguist, who is ultimately called to testify about language. Similarly, most of the times these techniques require plenty of data for training and testing, which is not the standard scenario in forensic linguistics. All of these aspects of computational authorship profiling therefore make these computational techniques very good for applications where

the objective is a fast scrutiny of large data sets, for example in marketing applications, but not always useful for the typical scenario of a forensic linguist being asked by the police about the most likely profile of the author of a one page threatening letter.

The present article argues that the development of a method of *forensic* authorship profiling for anonymous written texts can only come from research in two directions: (1) the accumulation of knowledge and understanding of stylistic variation across social factors, and (2) the verification that these patterns are also found in the register of the disputed document to be profiled. The first direction addresses the need for established linguistic theory and knowledge to be applied to forensic scenarios. The second direction addresses a fact often ignored by computational research in authorship profiling, that is, the pervasive effect of register variation on language (Biber, 1995, 2012; Biber and Conrad, 2009).

This article reports on an experiment on English data aimed at identifying which stylistic patterns previously found in other studies can be used for profiling three demographic characteristics (gender, age, and social class) in a situation similar to the typical forensic linguistic scenario of an anonymous short letter.

Literature review

The pre-requisite to perform a forensic linguistic task such as profiling in a linguistically-informed way is to first carry out a complete survey of what is known about language variation and the social factors of interest. This literature review constitutes a survey of key research that could inform forensic authorship profiling for the three social factors considered: gender, age, and social class. Other social factors, such as ethnicity, could also be considered, but these three are a good starting point, given their potential investigative value as well as the existence of a large amount of linguistic research on stylistic variation associated with them. The literature review focuses only on those studies that can be used for the typical forensic linguistic scenario of the profiling of the style of an anonymous written text. The present work is not concerned with studies that looked at alternations such as *was* for *were* or *innit* as a tag question, as these features are the type of features involved in a Type 1 analysis. Instead, this review focuses on the established patterns of variation that have been found to distinguish the social groups considered in terms of, for example, the use of nouns as opposed to verbs, clausal patterns, and other lexicogrammatical features that are pervasive and therefore that are always found in any text, considering the text itself as a unit of analysis.

Gender

The notion and definition of the concept of gender is not trivial but despite these problems the profiling of someone's gender is a question that can be asked to forensic linguists. Although it has been often useful to draw a distinction between the socio-cultural cline of *gender* and a biological binary *sex*, there is evidence suggesting that in reality none of these constructs is binary (Bing and Bergvall, 1998). The tension in profiling work is that whereas law enforcement are interested in certain biological correlates of gender, the clues that can be found in language are more likely to reveal the socio-cultural gender of the author, which is a continuum as well as subject to variation depending on extra-linguistic context (Carothers and Reis, 2013). These issues have not been addressed extensively in stylistic research that involved gender and the research

reviewed below thus significantly simplifies the nature of this dependent variable, reducing it to a division between biological *men* and biological *women*. Despite this issue, this research is the only starting point for work on gender profiling at this stage.

The most important pattern identified by previous studies of stylistic variation and gender is in the continuum between *nominal vs. clausal style*, the former being more typical of men and the latter more typical of women. The nominal end of the continuum is more often characterised by use of features such as nouns, adjectives, prepositions, and, generally, complex noun phrases, whereas the clausal end is characterised by the use of features such as verbs, adverbs, and simple noun phrases constituted by pronouns only. This pattern has been extensively found in a large number of studies at different times and in several registers and the literature thus suggests that this is a pervasive effect, even though the reported effect sizes have been relatively small. This pattern has been found in structured sociolinguistic interviews (Poole, 1979, N = 96), casual conversations (Rayson *et al.*, 1997), personal letters (Biber *et al.*, 1998, N = 80), and large corpora of formal/informal and fiction/non-fiction written texts (Koppel *et al.*, 2002; Argamon *et al.*, 2003; Schler *et al.*, 2006; Newman *et al.*, 2008). A gender effect on the frequency of nouns and pronouns has also been observed diachronically by Säily *et al.* (2001, N = 660) in a corpus of letters dating from 1415 to 1681.

After analysing speech data from 80 participants and finding a similar effect, Heylighen and Dewaele (2002) have proposed that this pattern could be due to the level of formality, where formality indicates the level of mathematical preciseness of a text as opposed to its dependence on the extra-linguistic context. They introduce an index to measure formality defined as follows

$$F = \frac{(R_{nouns} + R_{adj} + R_{prep} + R_{art}) - (R_{pro} + R_{verbs} + R_{adv} + R_{interj}) + 100}{2}$$

where the first bracket contains the relative frequencies of the nominal/formal elements (nouns, adjectives, prepositions, and articles) and the second bracket contains the relative frequencies of the clausal/contextual elements (pronouns, verbs, adverbs, and interjections).

Despite this attempt, the literature reveals that there is far more advancement in the description as opposed to the explanation for this pattern, especially since a clear definition of *gender* is still lacking. It has been proposed in the past that the sociolinguistic effects of gender could have both biological and social explanations (Chambers, 1992) and certain elements of the patterns described above have indeed been given a psychological explanation by social psychologists who have found that increased pronoun use correlates with gender in the direction described and with a tendency for neuroticism, which is also more common among women (Pennebaker *et al.*, 2003; Rude *et al.*, 2004). In a very small pilot study of only two subjects Pennebaker *et al.* (2004) found that an increase in testosterone levels increases the level of nominal style employed. On the other hand, other plausible explanations for this gender effect can be found in the tendency for these two genders to engage with different registers (Herring and Paolillo, 2006) and in the network of relationships that they therefore establish (Bamman *et al.*, 2014), and

thus, ultimately in the different *communities of practice* that the different genders on average engage with (Eckert and McConnell-Ginet, 1992).

Age

Although the concept of ageing would intuitively seem relatively unproblematic, from a linguistic point of view it is indeed much more multi-faceted. Statistically it is convenient to reduce age to a number as has been done in several studies but this measure of biological age might not be the best predictor of linguistic variation, as *social age*, as opposed to biological age, is more likely to affect language (Eckert, 1998). Social age is marked by a series of socially-recognised landmark events in life, such as certain birthdays, marriage, entering the job market, etc., which require different linguistic varieties and which offer different registers that can and sometimes must be learned. If this is correct, then profiling age has the same tension seen above for gender: whereas law enforcement is mostly interested in biological age, linguistic variation can only reveal clues as to the social age of a person, which is only a proxy for biological age.

The most well established pattern of stylistic variation that correlates with age is the negative relationship between syntactic complexity and ageing, which has been discovered in psycholinguistics. Analysing experimental data and diary entries, Kemper (1987) discovered that as people age there is a tendency to abandon complex clausal syntax, measured by average number of clauses per sentences, and in particular left-branching complexity. The explanation that they proposed for this pattern is that it is an effect of working memory, which decreases with age and especially in situations of dementia or Alzheimer's disease. This effect of age on syntax was found in several other studies of different and large subject groups and different registers, such as oral interviews and written essays (Kemper *et al.*, 1989, N = 108), descriptive essays (Bromley, 1991, N = 240; Rabaglia and Salthouse, 2011, N = 900), speech samples (Kemper and Sumner, 2001, N = 200), and in the famous Nun Study, in which autobiographic texts of a group of 150 nuns were analysed from 1930 until 1996 (Kemper *et al.*, 2001).

Interestingly, some of these studies, such as Kemper and Sumner (2001) and Rabaglia and Salthouse (2011), have also noticed an increase in vocabulary variety with old age, measured via type-token ratio or average word length. This effect of ageing on lexical richness is in line with the recurrent finding that ageing plays a role in the nominal vs. clausal style pattern already observed for gender. For example, Pennebaker and Stone's (2003) study of emotional disclosure essays and interviews on more than 3,000 participants found generally an increase in frequencies of determiners and prepositions and a decrease in frequency of pronouns with ageing. Likewise, Schler's *et al.* (2006) analysis of blogs written by almost 40,000 writers also found that as people age they tend to adopt a more informational/nominal style.

In sum, although ageing seems to be correlated with loss in syntactic complexity, another form of complexity based on nominal structures and lexical richness seems to replace it. This is consistent with another explanation for this effect given by Kemper *et al.* (1989), who suggested that the decline in usage of complex syntactic forms might be due to older people becoming more familiar with better ways of conveying meaning that do not involve unnecessarily complicated structures, relying more on refined vocabulary. This explanation is consistent with the effect found regarding nominal style, as this style is more characteristic of literate registers, such as academic and scientific registers (Biber, 1988), which take time and experience to be acquired.

Social class

Although years of research in variationist sociolinguistics have found that social class is one of the most important predictors of language use, authorship profiling so far has not devoted much research to this factor. The problem with social class is that it is a controversial and difficult factor to quantify, a controversy made worse by the virtual disengagement between linguistics and sociological literature (Ash, 2002). Very rarely do linguistic studies adopt the same definition of social class and yet this construct very often shows effects of large magnitude. Most of the indexes used are based on occupation, but they tend to include other aspects, such as level of education, income, household, and parents' backgrounds. Despite the problems and controversies and general lack of research in computational authorship profiling, these factors are typically useful information for investigators in a case involving an anonymous text and they are therefore a necessary inclusion in the practice of forensic authorship profiling.

Previous literature has found that overall higher classes are more competent in the use of complex syntax due to their more frequent exposure to this kind of input. This pattern is very well established, with studies that found effects on various pools of subjects across decades. Syntactic complexity, measured through average sentence length or number of dependent clauses per sentence, has been found to be correlated with class by Loban (1967, N = 211) in both oral and written texts, Poole (1976, N = 80) in life-forecast essays, Johnston (1977, N = 36) in experimental elicited narratives, Poole (1979, N = 96) in structured interviews, Labov and Auger (1993, N = 10) in sociolinguistic interviews, and it was also found in Kemper *et al.*'s (1989) and Mitzner and Kemper's (2003) studies on syntax and ageing to be a good predictor of level of education.

A measure often employed to study complexity and its relationship to social class and level of education is the complexity of *t-units*, where a t-unit is defined as an independent clause with all its dependent clauses. Average t-unit length or the number of clauses per t-unit are therefore better measures of complexity than average sentence length, as sentences are instead orthographic units. Several studies have found that the management of t-units and especially the way they are punctuated is characteristic of certain levels of education and class, with both the complexity and the ratio of t-units per sentences being a good proxy to the degree of competence with standard punctuation (Loban, 1967; Hunt, 1971, 1983).

Similarly to gender and age, social class seems to participate in the nominal vs. clausal pattern, with the higher social class being more familiar and thus more frequent users of the nominal style. Heylighen and Dewaele (1999) found that their measure of formality increased with the social status of their participants. Several studies found evidence for more frequent usage of nominal parts of speech in the discourse of higher social classes, such as uncommon adjectives in essays (Poole, 1976), subject noun phrases in elicited narration (Johnston, 1977, N = 36), nouns and adjectives in elicited narration (Hawkins, 1977, N = 263), or adjectives in sociolinguistic interviews (Macaulay, 2002, N = 45). This opposition between nominal vs. clausal style mirrors very well the distinction between restricted and elaborated codes made by Bernstein (1962), which he had already associated with social class.

Finally, several studies have found a relationship between lexical richness and social class or level of education. For example, very early studies such as Bernstein (1962, N = 106) or studies on readability measures (Kitson, 1921; Dubay, 2004) found that average

word length correlates with social status, a finding confirmed by Bromley (1991, N = 240) in descriptive essays, or by Berman (2008, N = 80) for narrative speech samples. Byrd (1993, N = 200), on the other hand, found that measures such as type-token ratio and the mean rarity score of a word were higher in various essays written by higher social classes, a finding confirmed by Mollet *et al.* (2010, N = 55) in student essays, in which they used a measure called Advanced Guiraud 1000, calculated using the following formula:

$$G = \frac{V - v}{\sqrt{N}}$$

where V indicates the total word types in a text, v indicates the *common* word types of the text, that is, the most common 1000 word types of a comparison corpus such as the British National Corpus, and N is the total number of word tokens in the text.

In sum, despite its controversial status, all the evidence points to a substantial effect of social class on language and this fact alone suggests that this social factor cannot and should not be ignored when profiling.

Methodology

In order to verify to what extent these patterns are found in *malicious forensic texts* (Nini, 2017), the ideal methodology would be to compile a corpus of such texts stratified by these three social factors. However, gathering such a corpus is an impossible enterprise as malicious forensic texts are rare on their own and even rarer are texts of this kind for which the demographics of the authors are reliably known. This study therefore adopts an experimental methodology which, despite the obvious drawback of not being based on naturally-occurring data, offers the key advantage of allowing greater control of the conditions. A common problem with corpus data for sociolinguistic studies, for example, is that it is not always possible to control very accurately the conditions under which data is produced and since register is a strong source of variation, this has the potential of skewing the results if it is not carefully isolated. With an experiment, on the other hand, the researcher can control the aspects of the situation that they wish and measure their effect on the factors.

Data

Ninety-six participants, all required to be native speakers of any variety of English, were recruited from different social backgrounds, such as university students, training police officers, members of a writing group for retired people, and homeless newspaper sellers. Most of the participants were from the UK and especially from England, with the exception of three participants from North America and one from Jamaica.

54% of the subjects declared their gender to be male and 46% to be female. For age, 37.5% of the participants were between 19 and 29 years old, 38.5% were between 30 and 50, and 24% were between 51 and 78. Finally, 55% of the participants did not have a university degree, while of the remaining 45%, 16% had an undergraduate degree and 29% had a postgraduate degree.

An index of social status was calculated using mainly the occupation of the subjects averaged over the occupation of their parents. A score from 1 (lower status) to 6 (higher

status) was assigned to each participant[1] using the classification of occupations adopted for the British National Corpus (McEnery, 2006: 27) in the following way:

- A - higher managerial, administrative or professional – Score 6
- B - intermediate managerial, administrative or professional – Score 5
- C1 - supervisory or clerical, and junior managerial, administrative or professional – Score 4
- C2 - skilled manual workers – Score 3
- D - semi- and unskilled manual workers – Score 2
- E - state pensioners or widows (no other earner), casual or lowest grade workers – Score 1

For students, only the average of their parents' score was considered.

A cross-tabulation of the factors revealed that the sample is very well balanced, with only a significant association between gender and age (binarized in two categories, *Older* and *Younger* at the median age of 38) (X^2 = 8.2, df = 1, p = 0.004), as there were more younger women than younger men overall. This skew could affect some of the results and it will be further discussed below. In addition, this analysis also revealed that the social class index is a good proxy to the education of the participants as the association between having a degree or not and belonging to the *Higher* or *Lower* class (based on the median index of 3.7) was significant (X^2 = 17.9, df = 1, p = 0.00002). The distribution of the participants in the corpus according to these categories can be seen in Table 1.

Higher		
	Male	*Female*
Older	13	9
Younger	10	21

Lower		
	Male	*Female*
Older	20	5
Younger	8	7

Table 1. Distribution of number of participants in the corpus across the three categories used in the study: Gender, Age, and Class.

The subjects were asked to fill in a questionnaire with details about themselves and to carry out a writing task in a computer lab in a university room and they were compensated with an expense and participation fee of £10. The subjects were asked to write three tasks that elicit three registers (see Appendix): (1) Task 1: a letter of complaint to a holiday agent asking for compensation; (2) Task 2: a letter to the Prime Minister of the United Kingdom to complain about the economic crisis and threatening not to vote for them again; (3) Task 3: a letter to a fictitious abusive employer threatening to damage their car if their behaviour does not change. The participants completed the three writings tasks in the same session and were not given any time constraints to finish the experiment. The simulated situation of these three texts was structured in particular to capture variation in the recipient: Task 1 is addressed to a company, Task 2 is addressed to a person of higher status and power that the participants do not personally

know, and Task 3 is addressed to a person of higher status that they personally know. In addition, the three tasks can all simulate potentially threatening letters to a company, a political figure, or an employer. The experimental tasks are similar for several situational parameters (Biber, 1994), such as being written with the possibility of editing, having no audience, not being specialised, etc. but they differ greatly in topic and, most importantly, in the level of knowledge between addressor and addressee. Since audience design has already been shown to be a very important predictor of linguistic variation (Bell, 1984), this difference is important and it is predicted to have a strong influence on the style of the participants.

Although the experiment consisted in eliciting texts that have been designed to capture scenarios as close as possible to real forensic cases, it is reasonable to argue that these are still elicited texts and therefore they may still be different from real authentic malicious forensic texts of this kind. To address this problem, Nini (2015) compared the experimental texts against a corpus of authentic malicious forensic texts described in Nini (2017) and found that the register of these experimental texts is almost indistinguishable from the register of real malicious forensic texts. The analysis was done by testing for statistically significant difference on 135 linguistic features that vary across registers, including the features of interest for profiling identified in this article. Only 13 out of 135 linguistic features were significantly different across the data sets but a qualitative scrutiny revealed that out of these 13 features only two were due to an experimental effect: contractions and proper nouns were used much more frequently in authentic texts than in fabricated texts for reasons attributable to differences between real and experimental conditions. However, since neither of these features seems to have a role to play in profiling, it can be concluded that the experimental texts are a good approximation to the register of real malicious forensic texts.

Features

The literature review has shown that there are consistent patterns of stylistic variation that correlate with the three social factors considered. Therefore, it is possible to make certain predictions about the relationship between language and the social factors that will be observed in the simulated malicious texts:

1. The nominal vs. clausal style would pattern in the following way: male/older/higher social class participants should exhibit a more nominal style than female/younger/lower social class participants. This stylistic cline can be measured using Heylighen and Dewaele's (2002) *F* score, which includes all the features explored in a number of other studies;
2. Higher class/older participants should use a richer vocabulary than lower class/younger participants. Vocabulary richness can be measured using several indices, such as average word length, type-token ratio, etc. For this study the Advance Guiraud 1000 score presented above was chosen as it is a more direct measure for estimating *extrinsic* vocabulary richness, or the rarity of the vocabulary used (Mollet *et al.*, 2010);
3. Higher class/younger participants should use a more complex clausal syntax than lower class/older participants. Sentence complexity can also be measured in several ways, for example simply using average sentence length. However, as noticed by Hunt (1971), a sentence is an orthographic unit and this is therefore

not ideal. For this reason, this study focuses on the number of clauses per *t-unit*, where a t-unit is an independent clause with all its dependent clauses. For this analysis, T-units were identified and segmented manually but the number of clauses was determined using a computer script that counted all the main verbs in the texts.

If these features, as predicted by previous studies, are unequally distributed across the social factors in these experimental texts that set out to simulate malicious forensic texts, then this is evidence that these principles of profiling can be used in real-life forensic cases involving similar registers. The differences were tested using non-parametric significance tests as the normality of the distributions is not assumed, using the Kruskal-Wallis tests for dependent variables with more than two categories and the Wilcoxon Rank Sum test for dependent variables with only two categories.

Results

As predicted, the most important finding of this study is the pervasive effect of register, as all the features considered exhibit substantial register variation. This was expected as it has been already demonstrated that register variation is the most important predictor of linguistic variation of the kind analysed in this study and in the studies reviewed. For this reason, all the results below are plotted using a mixture of two types of graph: boxplots showing differences across tasks with overlaid dotplots and point and range plots within these boxplots to show the differences for the social factors. This way of visualising the patterns also reflects the idea that these styles associated with the social factors are indeed ways of realising a particular register.

The second most important finding is the considerable importance of the nominal vs. clausal style pattern, which affects all social factors as predicted. Figure 1 shows how *F* has a very strong register effect ($p < 0.0001$) and how this effect is reflected in the social factors. Indeed, all the social factors have a significant effect in the predicted direction for *F* but only for Task 1 (Class, $p = 0.02$; Gender, $p = 0.04$; Age, $p = 0.02$), the more formal letter of complaint, while in Task 2 this difference is less strong and only significant for Class ($p = 0.02$) and in Task 3 all categories have cut median values around the median for the register and non-significant effects. For Class, it seems evident that the difference is mostly due to the Lower category, which includes all the participants with a score between 1 and 3. For age and gender, in Task 1 older participants and male participants both scored above the median for the register, as predicted.

Because all the factors interact, it is interesting to explore the pattern emerging from these factors when they are combined. Figure 2 plots the distribution of the *F* measure for cross-categories such as Class-Gender, Class-Age, and Age-Gender. For Task 1 and 2, the predictions are all correct: the top categories that include most of the texts that are far away from the median are Higher-Male, Higher-Older, and Older-Male while the categories that score far away from the median are the opposite, Lower-Female, Lower-Younger, and Younger-Female, with the categories in between scoring in the middle and very closely to the median for the group overall. All of these differences are relatively strong and mostly significant for Task 1 (Class-Gender, $p = 0.01$; Class-Age, $p = 0.0007$; Age-Gender, $p = 0.059$), less strong for Task 2 (and all non-significant, except for Class-Age, $p = 0.02$) and they are neutralised in Task 3, with none of the effects significant. It is important to note here that these plots show how the skew noted in the Methodology

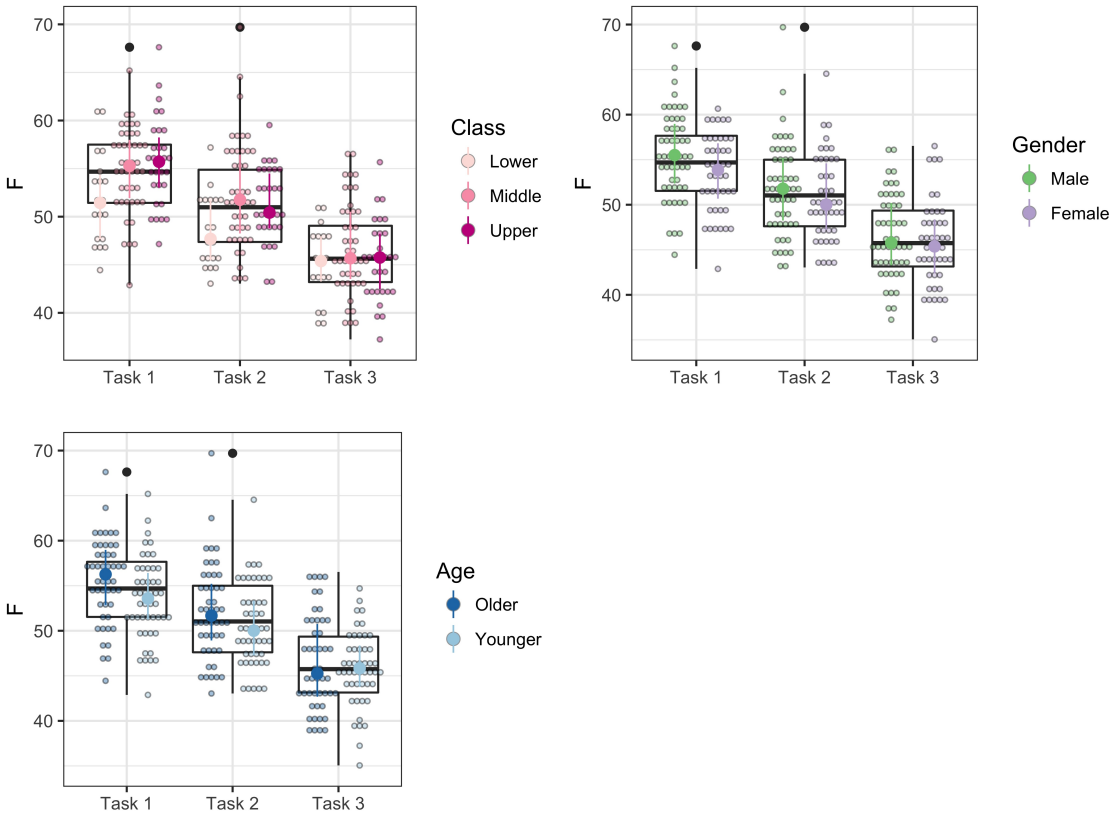

Figure 1. Boxplots showing the distribution of the *F* measure across Tasks. For each boxplot, a dotplot per social factor value is plotted containing information about the median (larger dot) and quartile (the range crossing the median dot).

section does not have a strong confounding effect in these results for *F*, as despite the relatively higher number of younger females in the sample, the predicted pattern is still observed.

Vocabulary richness measured through the Advanced Guiraud 1000 score also shows the predicted pattern, with a strong register effect ($p < 0.0001$) (Figure 3). Similarity to F, the predicted direction is observed here for all the tasks, except perhaps Task 3, with the category Higher-Older scoring the highest, the middle categories situated along the median for the registers, and the lowest category being Lower-Younger. However, in this case the effects are significant only for Task 2 ($p = 0.02$).

Finally, again the analysis of syntactic complexity using the measure of clauses per t-units confirms previous findings (Figure 4). For this measure of syntactic complexity, however, the register differences are far less accentuated, although still very significant ($p = 0.001$). The difference seems to involve mostly Task 2, which has a higher syntactic complexity overall than the other two registers.

In this case, the literature would predict that the highest scores for syntactic complexity would be obtained by the youngest members of higher social classes and the analysis reveals that this is the case, with a cline that follows the predictions. However, this effect is significant again only for Task 2 ($p = 0.03$), the register characterised by the highest median syntactic complexity overall.

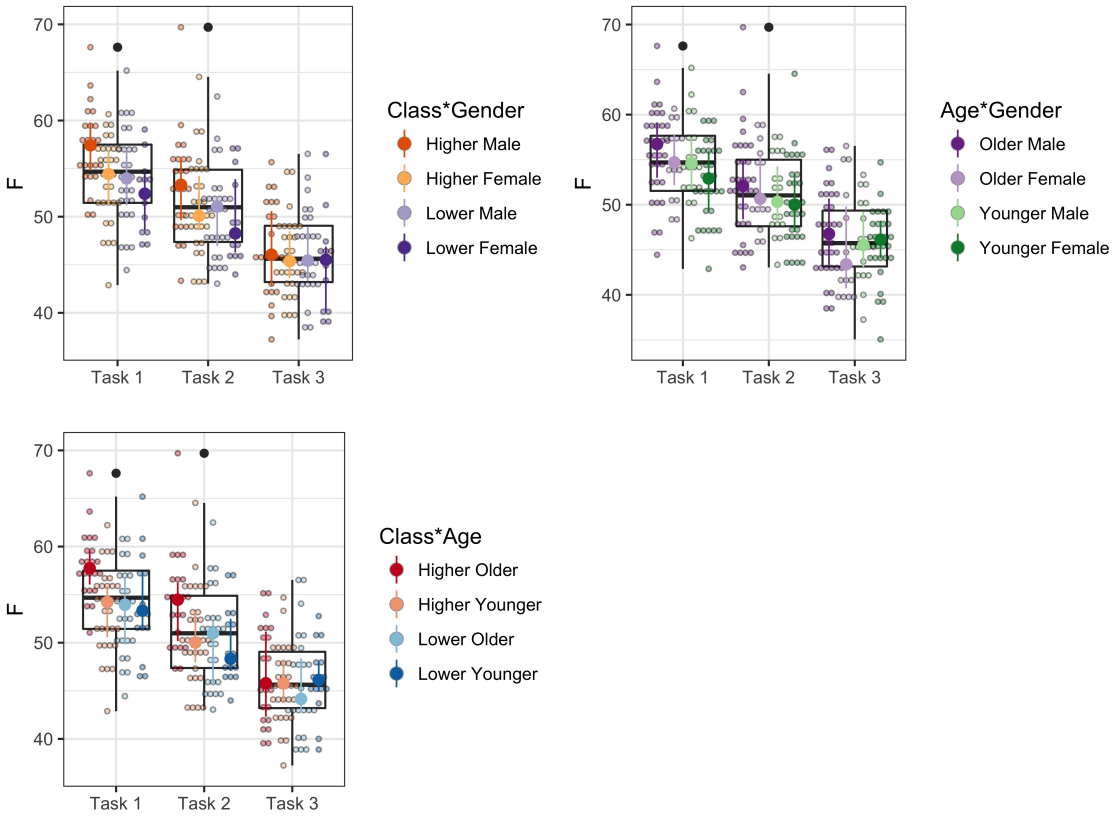

Figure 2. Boxplots showing the distribution of the *F* measure across Tasks for the social factors combined. For each boxplot, a dotplot per social factor value is plotted containing information about the median (larger dot) and quartile (the range crossing the median dot).

Discussion

These results firstly indicate that the nature of the linguistic features considered requires that an analysis of the register of the text in question is conducted before any profiling, as the register effect of these features is generally much stronger than any social factor effect. However, provided that this is done, the results reported in this paper suggest that the relationship between stylistic variation and social factors previously identified are generalisable to registers similar to malicious forensic texts. These findings also suggest that even though the effects seem stable and unlikely to reverse direction, they do not necessarily appear in all registers. Therefore, although it would be very unlikely to find, for example, younger women to have a higher *F* score than older men in any register, it is possible that the predicted effect is neutralised by register effects. In other words, these findings suggest that register gives the space for stylistic variation of this kind to occur, as can be seen in the analysis of syntactic variation, which presents a social effect only for Task 2 where the amount of clausal complexity is overall higher.

Because of this strong register effect, it is fair to conclude that it is unlikely that any of these effects are exclusively the results of biological or psychological factors such as working memory. For example, if syntactic complexity decreased only because of a decrease in working memory capacity, then the same effects observed for Task 1 should be observed in Task 3. Explanations should instead be sought in particular in the reasons

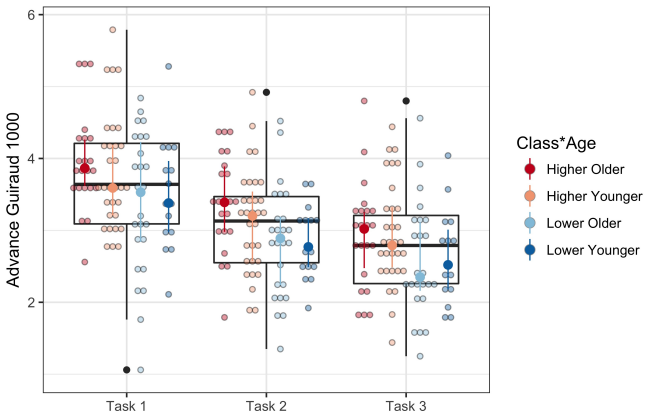

Figure 3. Boxplots showing the distribution of the Advance Guiraud 1000 score across Tasks for Social class and Age. For each boxplot, a dotplot per social factor value is plotted containing information about the median (larger dot) and quartile (the range crossing the median dot).

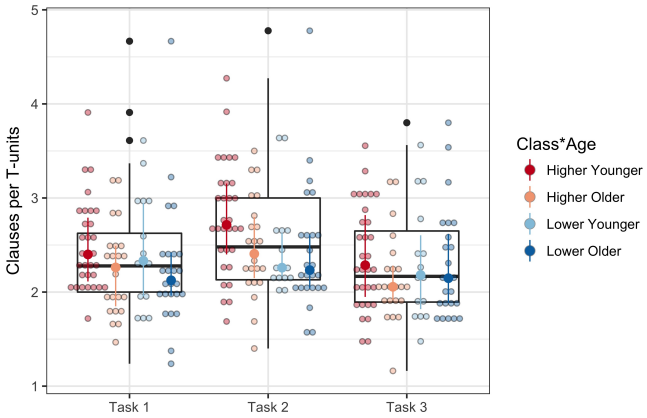

Figure 4. Boxplots showing the distribution of the number of clauses per T-units across Tasks for Social class and Age. For each boxplot, a dotplot per social factor value is plotted containing information about the median (larger dot) and quartile (the range crossing the median dot)

why these social categories employed these styles and on the nature of the relationship between styles and registers.

To explain the relationship between stylistic variation and social factors, let us consider the case of *F*, the most important linguistic feature of this study representing the opposition between nominal vs. clausal style. This stylistic contrast has been found across several studies in English and other languages and has been named in different ways. Heylighen and Dewaele (1999) expressed this contrast in terms of reliance on context and formality, while Biber's (1988) multidimensional study named this stylistic contrast functionally as the opposition between informational and involved discourse. More recently, Biber (2014) has renamed this opposition as the contrast between clausal vs. phrasal discourse. The findings of this study completely support previous findings: the *F* measure increases as the personal knowledge between interactants decreases because

a higher degree of distance between addressor and addresse(s) requires less reliance on context and thus a more pervasive adoption of nominal features of elaboration.

Crucially, these register differences for *F* and for the other features might also be responsible for the social differences observed in this and other studies, as explained by the *register axiom*. Finegan and Biber (2001: 265) define the register axiom as follows:

> If a linguistic feature is distributed across social groups and communicative situations or registers, then the social group with greater access to the situations and registers in which the features occur more frequently will exhibit more frequent use of those features in their social dialects.

In Systemic Functional Linguistics, this is expressed in the theory of *codal variation* or *semantic variation*, for which social groups differ in terms of the meanings that they make as well as the linguistic features they use and that this difference is due to the different degree of access that social groups have to certain registers (Hasan and Cloran, 1990; Hasan, 1996, 2009). This theory can help in explaining, at least partially, the effects observed, and in particular the results regarding the clausal vs. nominal style, as the nominal style is more frequently encountered in written formal writings and only members of the higher classes who work in occupations in which they often encounter this nominal style can therefore develop competence with it. This theory can also help explaining the neutralisation effect of register: Task 3, which does not require a nominal style, does not lead to social differences because even the higher social classes, who are capable of using the nominal style, still choose to use the clausal style as it is the most appropriate for the context. Suggestions along similar lines have been proposed by Bernstein (1962), who proposed that there are two codes of expression, the *restricted* and the *elaborated* code, and that social inequality arises as only certain occupations have access to both codes. If the nominal style can be compared to the elaborated code, then these results are compatible with his theory and provide a linguistically justified explanation for the social effect that can inform the profiling task.

However, although richer vocabulary and familiarity with the nominal style for certain participants in certain occupations are both explainable with their greater familiarity with certain registers, other effects cannot be easily explained using only the register axiom. For example, the difference in the *F* score found for males vs. females, although modest, cannot be explained by the unequal gender access to certain occupations, since occupation and social class were controlled in this experiment. It seems that even with equal access to certain registers, men tend to score on average higher on *F* than women and therefore there must be other factors at play. Similarly, the psycholinguistic literature has demonstrated that older speakers tend to use less complex syntax partially because of decrease in working memory and this effect cannot be completely discounted. These considerations lead to the more general conclusion that authorship profiling might not be an exclusive *sociolinguistic* phenomenon and it would therefore be partially erroneous or misleading to refer to the task of authorship profiling as simply *sociolinguistic profiling*, as certain linguistic patterns might have explanations outside of the field of sociolinguistics, for example in psycholinguistics.

The last consideration is about the importance of taking a measure of social class or occupation into account, as previous linguistics literature and this study show how this social factor has the largest effect on language. Virtually no computational authorship profiling research has been devoted to profiling social class or occupation, and this is

problematic as these results show how much an impact this social factor has, even if the goal is the profiling of other demographics.

Conclusions: how to develop forensic authorship profiling?

In sum, it seems very difficult at the present stage for an automatic system to be able to untangle all of the factors that this study has outlined, from the importance of register to the interaction of the social factors, especially if the text to be profiled is very short, as is common in forensic linguistics. Carrying out profiling for forensic purposes means, in essence, estimating the most likely demographics of the author of one of the dots in any of the four figures above. As an example, let us assume a questioned text has been analysed and its *F* score is 45. Looking at Figures 1 and 2 it is evident that whereas a score of 45 is completely the norm for a text like Task 3, it is definitely outside the norm for a text like Task 1, for which this score is very unlikely and only found in the lower classes. The understanding of the register is therefore a precursory step for profiling. However, even an analysis of the register does not substantially help in the majority of cases. The clouds of points in those graphs makes it evident that there is a great degree of overlap between the categories and, consequently, not very much discriminatory potential. Profiling of the general demographics is therefore a very difficult task, which might be possible only in certain extreme circumstances, such as when the questioned texts behave in ways that are substantially outside the norm. For example, the results of this study show how an *F* score of 60 for Task 1 is very unlikely for the average Lower-Female but typical for the average Higher-Male.

The crucial step for carrying out profiling right now thus seems to be the identification of deviation from a norm. For example, although it now seems established that higher social classes/men/older individuals use a more nominal style than lower social classes/women/younger individuals, what *more* and *less* mean depends on the register of the questioned text, which should therefore be analysed before carrying out profiling. My proposal for an algorithm for the forensic authorship profiling of writing style based on these considerations is therefore as follows:

1. Study the extra-linguistic situation of the questioned text, for example using Biber's (1994) Situational Parameters;
2. Collect and analyse a corpus with comparable situational parameters to establish the norm for the linguistic features that will be analysed, the set of which should be based on previously established literature on stylistic variation. If possible, the corpus should contain texts written by a stratified sample of the population to verify that the previously established stylistics patterns are present and whether they follow the predicted direction and to what extent;
3. Check the position of the disputed text in the register space given by the comparison corpus, similarly to the graphs presented above, so that the position of the text in relation to the distribution for the register can be assessed;
4. Bearing in mind previous literature, of which this article is an initial survey, compare the linguistic behaviour of the disputed text against the norm;
5. Very importantly, the meaning of the numbers should not be ignored, especially for short texts. Knowledge from previous literature is useful because it provides an explanation for the linguistic patterns that we observe but only if the linguistic patterns can be explained by the same principles can these be used to infer the characteristics of the anonymous author.

The most challenging component of this algorithm is probably step (2), as it might be difficult or impossible to collect a stratified sample of certain registers. However, this is what core research in *forensic* authorship profiling should do: focus on expanding on the present work so that a forensic linguist does not have to collect an *ad hoc* corpus for every case and can therefore use previous studies for direct comparison. For example, the study reported here could be used as a baseline for forensic work on a questioned text with situational parameters similar to one of the three Tasks, even though replications of this study are also, of course, highly encouraged.

For the future, two items are particularly urgent in the agenda: (1) to increase understanding of the social factors that are profiled, and (2) to develop new computational techniques that are aware of these issues and that include linguistic theory.

The first point concerns the issues raised in the literature reviews above regarding the definition of the three social factors, gender, age, and social class. It is unquestionable that these categories cannot be simply defined in the way that has been used in previous studies and, consequently, in this present study. However, there is a problematic tension between the requirements of law enforcement and what an analysis of language can reveal. In all likelihood linguistic analyses can only profile social factors that are proxies to the type of social information that law enforcement needs and future research into *forensic* authorship profiling should address this tension. For example, more studies should focus on untangling the elements of gender that correlate with stylistic variation, so that it is clear, for instance, what the *F* score is actually measuring. Equally, studies are needed to verify whether biological age is indeed a proxy to social age in terms of stylistic variation. This knowledge can inform the type of inference that a forensic linguist can make when faced with a profiling problem.

The second point concerns the direction of research and the collaboration between computer scientists and linguists. There is no doubt that more sophistication in the analysis can help with the issues outlined in this article and this level of sophistication can certainly only come from the fields of computer science and computational statistics. However, the research in these fields should be guided both by the needs and, more importantly, by the previous knowledge already available in the fields of enquiry in which these statistical and computational techniques are applied, that is, linguistics. This collaboration can ensure that sophistication of method is paired with a high degree of interpretability and that it is also contextualised within the field of linguistics. It is likely that a method based on machine learning, such as that of Argamon *et al.* (2009), if applied to the present data sets would still return good accuracy rates and, if trained with appropriate awareness of register issues, even achieve better performance. However, it is still debatable to what extent these results would be useful in a forensic context without a proper linguistic interpretation.

The understanding of the underlying linguistic patterns responsible for the predictions is a pre-requisite for *forensic* authorship profiling because, ultimately, the evidence analysed is linguistic and not statistical. Therefore, although computational methods can and should be employed to aid the analysis, this must not be done at the expense of the underlying linguistic explanations, which should remain the primary focus within forensic linguistics.

In conclusion, because of what is at stake in a forensic setting, authorship profiling can be developed into *forensic* authorship profiling only when linguistics and computer science work side by side keeping the focus not on techniques but on linguistic explanations, theories, and knowledge, with particular attention to the forensic context.

Notes

[1]With the exception of three participants for whom it was not possible to obtain occupation information about their parents.

References

Argamon, S., Koppel, M., Fine, J. and Shimoni, A. R. (2003). Gender, genre, and writing style in formal written texts. *Text - Interdisciplinary Journal for the Study of Discourse*, 23(3), 321–346.

Argamon, S., Koppel, M., Pennebaker, J. W. and Schler, J. (2009). Automatically profiling the author of an anonymous text. *Communications of the ACM*, 52(2), 119–123.

Ash, S. (2002). Social class. In J. K. Chambers, P. Trudgill and N. Schilling-Estes, Eds., *The Handbook of Language Variation and Change*. Malden, MA; Oxford: Blackwell Publishers, 402–423.

Bamman, D., Eisenstein, J. and Schnoebelen, T. (2014). Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2), 135–160.

Bell, A. (1984). Language style as audience design. *Language in Society*, 13, 145–204.

Berman, R. (2008). The psycholinguistics of developing text construction. *Journal of child language*, 35(4), 735–71.

Bernstein, B. (1962). Linguistic codes, hesitation phenomena and intelligence. *Language and Speech*, 5(4), 221–240.

Biber, D. (1988). *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D. (1994). Register and social dialect variation: An integrated approach. In D. Biber and E. Finegan, Eds., *Sociolinguistic Perspectives on Register*. Oxford: Oxford University Press, 315–347.

Biber, D. (1995). *Dimensions of Register Variation: a Cross-Linguistic Comparison*. Cambridge; New York: Cambridge University Press.

Biber, D. (2012). Register as a predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory*, 8(1).

Biber, D. (2014). Using multi-dimensional analysis to explore cross-linguistic universals of register variation. *Languages in Contrast*, 14(1), 7–34.

Biber, D. and Conrad, S. (2009). *Register, Genre, and Style*. Cambridge; New York: Cambridge University Press.

Biber, D., Conrad, S. and Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Bing, J. M. and Bergvall, V. L. (1998). The question of questions: beyond binary thinking. In J. Coates and P. Pichler, Eds., *Language and Gender: a Reader*. Chichester, West Sussex, U.K.; Malden, MA: Wiley-Blackwell, 495–511.

Bromley, D. B. (1991). Aspects of written language production over adult life. *Psychology and Aging*, 6(2), 296–308.

Byrd, M. (1993). Adult age differences in the ability to write prose passages. *Educational Gerontology: An International Quarterly*, 19, 375–396.

Carothers, B. J. and Reis, H. T. (2013). Men and women are from Earth: Examining the latent structure of gender. *Journal of Personality and Social Psychology*, 104(2), 385–407.

Chambers, J. K. (1992). Linguistic correlates of gender and sex. *English World-Wide*, 13(2), 173–218.

Dubay, W. H. (2004). *The Principles of Readability*. Costa Mesa: Impact Information.

Eckert, P. (1998). Age as a sociolinguistic variable. In F. Coulmas, Ed., *The Handbook of Sociolinguistics*. Oxford, UK; Cambridge, Mass: Blackwell Publishers, 151–167.

Eckert, P. and McConnell-Ginet, S. (1992). Think practically and look locally: Language and gender as community-based practice. *Annual Review of Anthropology*, 21, 461–490.

Finegan, E. and Biber, D. (2001). Register variation and social dialect variation: The register axiom. In P. Eckert and J. R. Rickford, Eds., *Style and Sociolinguistic Variation*. Cambridge: Cambridge University Press, 235–267.

Grant, T. (2008). Approaching questions in forensic authorship analysis. In J. Gibbons and M. T. Turell, Eds., *Dimensions of Forensic Linguistics*. Amsterdam: John Benjamins Publishing Company, 215–231.

Hasan, R. (1996). Ways of saying: ways of meaning. In C. Cloran, D. Butt and G. Williams, Eds., *Ways of Saying, Ways of Meaning: Selected Papers of Ruqaiya Hasan*. London: Cassell, 191–242.

Hasan, R. (2009). Wanted: a theory for integrated sociolinguistics. In J. Webster, Ed., *The Collected Works of Ruqaiya Hasan Vol. 2: Semantic Variation: Meaning in Society and in Sociolinguistics*. London: Equinox, 5–40.

Hasan, R. and Cloran, C. (1990). A sociolinguistic interpretation of everyday talk between mothers and children. In M. A. K. Halliday, J. Gibbons and H. Nicholas, Eds., *Learning, Keeping, and Using Language. Volume 1*. Amsterdam; Philadelphia: John Benjamins Publishing, 67–100.

Hawkins, P. R. (1977). *Social Class, the Nominal Group and Verbal Strategies*. London: Routledge and Kegan Paul.

Herring, S. C. and Paolillo, J. C. (2006). Gender and genre variation in weblogs. *Journal of Sociolinguistics*, 10(4), 439–459.

Heylighen, F. and Dewaele, J. (1999). *Variation in the contextuality of language: an empirical measure, Center "Leo Apostel"*. Brussels: Free University of Brussels.

Heylighen, F. and Dewaele, J. (2002). Variation in the contextuality of language: an empirical measure. *Foundations of Science*, 1–27.

Hunt, K. (1971). Teaching syntactic maturity. In *Applications of Linguistics: Selected Papers of the Second International Congress of Applied Linguistics*, 287–301, Cambridge: Cambridge University Press.

Hunt, K. (1983). Sentence combining and the teaching of writing. In M. Martlew, Ed., *The Psychology of Written Language: Developmental and Educational Perspectives*. New York: John Wiley, 99–125.

Johnston, R. (1977). Social class and grammatical development: A comparison of the speech of five year olds from middle and working class backgrounds. *Language and Speech. SAGE Publications*, 20(4), 317.

Kemper, S. (1987). Life-span changes in syntactic complexity. *Journal of Gerontology*, 42(3), 323–328.

Kemper, S., Greiner, L. H., Marquis, J. G., Prenovost, K. and Mitzner, T. L. (2001). Language decline across the life span: Findings from the Nun Study. *Psychology and Aging*, 16(2), 227–39.

Kemper, S., Kynette, D., Rash, S., O'Brien, K. and Sprott, R. (1989). Life-span changes to adults' language: Effects of memory and genre. *Applied Psycholinguistics*, 10(01), 49–66.

Kemper, S. and Sumner, A. (2001). The structure of verbal abilities in young and older adults. *Psychology and Aging*, 16(2), 312–322.

Kitson, H. D. (1921). *The Mind of the Buyer*. New York: MacMillan.

Koppel, M., Argamon, S. and Shimoni, A. R. (2002). Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4), 401–412.

Labov, W. and Auger, J. (1993). The effect of normal aging on discourse: A sociolinguistic approach. In H. H. Brownell and Y. Joanette, Eds., *Narrative Discourse in Neurologically Impaired and Normal Aging Adults*. San Diego, California: Singular Pub Group, 115–135.

Leonard, R. (2005). Forensic Linguistics: Applying the Scientific Principles of Language Analysis to Issues of the law. *The International Journal of the Humanities*, 3.

Loban, W. (1967). *Language Ability - Grades Ten, Eleven, and Twelve. Final Report*. Rapport interne, Berkeley.

Macaulay, R. (2002). Extremely interesting, very interesting, or only quite interesting? Adverbs and social class. *Journal of Sociolinguistics*, 6(3), 398–417.

McEnery, T. (2006). *Swearing in English*. London: Routledge.

Mitzner, T. and Kemper, S. (2003). Oral and written language in late adulthood: Findings from the Nun Study. *Experimental Aging Research*, 29, 457–474.

Mollet, E., Wray, A., Fitzpatrick, T., Wray, N. R. and Wright, M. J. (2010). Choosing the best tools for comparative analyses of texts. *International Journal of Corpus Linguistics*, 15(4), 429–473.

Newman, L. M., Groom, C. J., Handelman, L. D. and Pennebaker, J. W. (2008). Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes*, 45(3), 211–236.

Nini, A. (2015). *Authorship Profiling in a Forensic Context*. Phd thesis, Aston University, UK.

Nini, A. (2017). Register variation in malicious forensic texts. *International Journal of Speech Language and the Law*, 24(1), 99–126.

Pennebaker, J. W., Groom, C. J., Loew, D. and Dabbs, J. M. (2004). Testosterone as a social inhibitor: Two case studies of the effect of testosterone treatment on language. *Journal of Abnormal Psychology*, 113(1), 172–175.

Pennebaker, J. W., Mehl, M. R. and Niederhoffer, K. G. (2003). Psychological aspects of natural language use: our words, our selves. *Annual Review of Psychology*, 54, 547–77.

Pennebaker, J. W. and Stone, L. D. (2003). Words of wisdom: Language use over the life span. *Journal of Personality and Social Psychology*, 85(2), 291–301.

Poole, M. (1976). *Social Class and Language Utilization at the Tertiary Level*. St. Lucia, Q: University of Queensland Press.

Poole, M. E. (1979). Social-class, sex and linguistic coding. *Language and Speech*, 22, 49–67.

Rabaglia, C. and Salthouse, T. (2011). Natural and constrained language production as a function of age and cognitive abilities. *Language and Cognitive Processes*, April 2013, 37–41.

Rayson, P., Leech, G. and Hodges, M. (1997). Social differentiation in the use of English vocabulary: some analyses of the conversational component of the British National Corpus. *International Journal of Corpus Linguistics*, 2(1), 133–152.

Rude, S., Gortner, E.-M. and Pennebaker, J. (2004). Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, 18(8), 1121–1133.

Säily, T., Siirtola, H. and Nevalainen, T. (2011). Variation in noun and pronoun frequencies in a sociohistorical corpus of English. *Literary and Linguistic Computing*, 26(2), 167–188.

Schler, J., Koppel, M., Argamon, S. and Pennebaker, J. (2006). Effects of age and gender on blogging. In *2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*, 199–205, Stanford, CA.

Shuy, R. (2014). *The Language of Murder Cases: Intentionality, Predisposition, and Voluntariness.* Oxford: Oxford University Press.

Appendix – The Experiment Tasks

Thank you for agreeing to participate in this experiment. The study is concerned with cases of interaction that are unfavourable or undesirable for the addressee.

The experiment consists of three tasks. For each, you will need to put yourself imaginatively in the situation that is described and then write a short text (at least 300 words) according to the guidelines provided.

The information you provide will be treated confidentially and will not be used for purposes other than the statistical measurement required for the present study.

SITUATION (1): Last year you bought a travel package from the FirstHoliday travel agency. Unfortunately, the holiday was totally unsatisfactory and you feel that it was not worth the price you paid. Indeed, you feel that the company should give you a refund.

TASK (1): Write a letter to the agency. You must not only express your feelings of dissatisfaction, but also describe how and why the situation made you very upset and angry. Warn them about possible legal action and ask for a partial refund of £500.

SITUATION (2): The economic crisis is making your life significantly more difficult. You feel frustrated that the coalition government is not addressing the issue as seriously as it deserves and you are worried that you might lose your job in the next few months if the planned cuts are not rescinded. You therefore think it is time to send a letter to them to make sure they understand that voters like you are unhappy and desperate.

TASK (2): Write an anonymous letter, signed as "A disappointed voter", to the Prime Minister showing your disappointment in how the government is managing the economic crisis. Express how the recession has hit you and that you are very angry that nothing has been done to prevent the situation. Make it very clear that you won't vote for them again if they don't change policies.

SITUATION (3): You are an employee of a company where you have been working for a long time. You have a newly appointed boss who is extremely abusive to you and to your colleagues and apparently does not value your work. To scare your boss, you are planning to make him think that if he does not change his unreasonable behaviour, someone will damage his car.

TASK (3): Write an anonymous letter, signed as "An angry employee", where you express your thoughts and feelings about his abusive behaviour. As well as expressing your views, scare your boss by using one of the following options for each category:

(a) car parts to be damaged: bodywork mirrors – tyres – lights
(b) object used to damage: baseball bat – jack – nail – spray paint
(c) time: early morning – lunch break – night

The creation of Base Rate Knowledge of linguistic variables and the implementation of likelihood ratios to authorship attribution in forensic text comparison

Sheila Queralt

SQ – Lingüistas Forenses, Spain

Abstract. *This article contributes to the research challenges that Forensic Linguistics faces in the 21st century – to compare texts of unknown authorship with the same reliability as other disciplines that consider forensic evidence. This research implements advanced statistical techniques within the field of forensic text comparison that improve the reliability of linguistic evidence furnished in Court and assess its significance. The first part of the analysis creates a Base Rate Knowledge for some of the most relevant linguistic variables in Peninsular Spanish texts. The second part applies statistical tests to variables with discriminatory potential to identify the samples of the authors and also assesses the reliability of the results in a posteriori classification. The implementation of the likelihood-ratio framework in the third part improves the reliability of linguistic evidence provided in court and offers probabilistic results to assist not only the judge and jury but also the linguistic expert in order to carry out more rigorous testing and extensive performance analysis of the data.*

Keywords: *Forensic text comparison, Authorship Analysis, Idiolect, Multivariate methods, Likelihood ratios.*

Resumo. *Este artigo contribui para os desafios da investigação enfrentados pela Linguística Forense no século XXI, de modo a comparar textos de autoria desconhecida com a mesma fiabilidade que outras disciplinas que consideram a prova forense. Este estudo implementa técnicas estatísticas avançadas na área da comparação de textos forenses para aumentar a fiabilidade da prova linguística fornecida em Tribunal e para avaliar a sua significância. A primeira parte da análise cria uma base de referência para algumas das variáveis linguísticas mais relevantes em textos de espanhol Peninsular. A segunda parte aplica testes estatísticos a variáveis com capacidade discriminatória para identificar as amostras dos autores, bem como avaliar a fiabilidade dos resultados em classificação a posteriori. A implementação de um quadro de razão de verosimilhança na terceira parte aumenta a fiabilidade da prova linguística fornedida em tribunal e oferece resultados probabilísticos para apoiar, não só o juiz e o júri, mas também o perito*

linguístico, de modo a realizar testes mais rigorosos e uma vasta análise do desempenho dos dados.

Palavras-chave: *Comparação de texto forense, Análise de autoria, Idioleto, Métodos multivariados, Razão de verosimilhança.*

Introduction

Over the last decades courts from several countries such as the United States, the United Kingdom or Spain have increasingly called on the expertise of linguists. The cases in which expert linguists give evidence can be diverse, from disputes about plagiarism, to trademarks, voice identification, linguistic profiling or authorship attribution cases. But the most frequent cases in forensic linguistics involve the comparison of an unknown sample (anonymous text) and a set of known texts from a suspect or several suspects. The estimation of the similarity between those two or more sources was traditionally approached by linguists using a verbal scale which may be based on estimations of probabilities or on opinion thresholds set by the expert (see for example Broeders 1999, Champod and Evett 2007 or Sjerps and Biesheuvel 2007). This traditional approach can be conceived to an extent as quite subjective considering that it is based on the linguistic expert's experience and may vary from expert to expert.

In the past, this traditional approach has been consigned to other forensic sciences that consider evidence such as DNA, fingerprints or handwriting. In parallel to the guidelines established, among other institutions, by the Committee on Identifying the Needs of the Forensic Sciences Community, which, for instance, states in its report that "a strong and reliable forensic science community is needed to maintain homeland security" (2009), therefore pointing towards the need of consolidating forensic techniques, the volume of forensic evidence and sophisticated forensic methods have increased over the last two decades. Consequently, multivariate and probabilistic methods have been developed in an attempt to evaluate the strength of the comparison of the quantifiable properties of known and unknown samples.

The most renowned probabilistic methodology across a broad spectrum of forensic sciences is the Likelihood-Ratio (henceforth LR) framework. In the last decade, research has proved the validity of LR models for assisting experts in forensic sciences to interpret evidence (Aitken and Taroni, 2004; Evett, 1998) and in the words of Fenton and Neil (2012: 2) expressing the "proper use of probabilistic reasoning has the potential to improve dramatically the efficiency and quality of the entire criminal justice system". Furthermore, the LR methodology fulfils the new needs of forensic individualization, applying transparent and testable procedures.

In the light of the aforesaid considerations, this article proposes the implementation of multivariate statistical methods and the LR framework for forensic text comparison through the analysis of linguistic variables. This methodology is implemented in threat texts written in Peninsular Spanish.

Methodological and theoretical framework

The concept of 'idiolect' has been the centre of some sociolinguistic variation studies such as Abercrombie (1969), Biber (1988), Biber (1995), Biber *et al.* (1998), Guy (1980) and also forensic linguistics studies, for instance, Queralt and Turell (2012), Cicres Bosch (2007), Gavaldà Ferré (2011), Spassova and Grant (2008), Spassova (2009) and Turell

(2010). The hypothesis of the existence of an idiolect makes it possible to establish a measure of idiolectal similitude to be able to state the probability of whether two linguistic samples have been produced by the same writer or not. This approach is widely accepted by the forensic linguistics community around the world as the approach to deal with the problem of questioned authorship. Nevertheless, the theory of idiolect is one of the long-standing and ongoing debates in the discipline. A number of scholars have identified practical issues that prevent this axiom from being demonstrated (e.g. Coulthard 2004: 432, Turell 2010: 217, Wright 2013: 46-47). And some have relied on alternative concepts to explain why forensic text comparison is possible, such as idiolectal style, consistency or pair-wise distinctiveness between authors (see, for instance, Turell 2010 and Grant 2010).

However, in this study the author wants to highlight that it is possible that every single person has a unique idiolect, but whether or not that is the case, it is surely true that people do develop a style and that each person's style is distinguishable from the styles of most other writers. As such, the more successful a method is in measuring the distance between the styles of different authors (even those of people with similar linguistic backgrounds), the more it should be viewed as a useful method.

In forensic text comparison, as in forensic voice comparison, the analysis of linguistic evidence does not consist only in describing the linguistic features that the unknown text contains. It also implies determining the degree of similarity between the writer's dependent features obtained from the unknown sample, and the writer's dependent features obtained from the known sample by the suspect (Gonzalez-Rodriguez *et al.*, 2006: 332).

A variety of different approaches have been developed within our discipline in the quest for quantifying the degree of similarity between samples such as relative frequency of functional or grammatical words (e.g. Burrows 1987 and Burrows 2003), word frequency distributions (e.g. Holmes 2003), vocabulary analysis (e.g. Coulthard 2004, Woolls and Coulthard 1998), and Part of Speech n-grams (e.g. Bel *et al.* 2012, Queralt *et al.* 2011, Queralt and Turell 2012, Spassova and Turell 2007, Turell 2004b and Turell 2004a); and also within other disciplines with more computational aspects, such as Juola (2006), Koppel *et al.* (2009) or Stamatatos (2009).

Nevertheless, quantifying the degree of similarity is not enough in forensic text comparison, one must also consider the rarity or the expectancy of those similar features compared to the relevant population. Coulthard and Johnson (2007) wonder "how can one measure the 'rarity' and therefore the evidential value of individual expressions" (p. 6). In order to calculate the degree of similarity and rarity between written samples one must estimate the population distribution – Base Rate Knowledge – of the relevant linguistic variables in a relevant population (Queralt, 2014: 43). These questions can be addressed by the use of these newly developed probabilistic methods, such as the Likelihood Ratio, which carries out rigorous empirical analyses. Unlike other kinds of evidence such as DNA profile data, forensic linguists deal with continuous and variable data and therefore the analysis has to consider two sources of variability: "the variability within the source (e.g., window) from which the measurements were made and the variability between the different possible sources (e.g., windows)." (Aitken and Taroni, 2004: 322).

In forensic linguistics, we use *inter-individual variation* to refer to the variability between writers and *intra-individual variation* for the variability within one writer. Intra-individual variation, variations across texts written by one author, is another intrinsic characteristic of linguistic data (see Labov 1972: 122, 127, 271-72, 319-25, Chambers 2009: 33-37 and Turell 1995: 20-22). Labov (1972: 208) states that "as far as we can see, there are no single-style speakers. Some informants show a much wider range of style shifting than others, but every speaker we have encountered shows a shift of some linguistic variables as the social context and topic change."

Intra-individual variations may occur in word choice, syntactic structures, grammatical patterns or in other linguistic levels and may be due to genre, time, social context, style, register or other external factors. According to the Saussurean view, the expert can handle intra-individual variation in two ways: on the one hand by treating idiosyncrasy as deviance and, on the other hand, by conceiving the linguistic individual as the set of strategic adaptations chosen from a closed set of conventional possibilities (Johnstone, 1996: 14).

Methodology

The world of forensic sciences is in continuous change due to the evolution of new technologies and the creation of more rigorous standards. Thus, in order to remain efficient and reliable, forensic sciences – in this particular case, forensic linguistics – need to adapt to these ongoing changes. This research intends to be viewed as a step forward in the direction the field should continue to evolve so as to increase its legitimacy as a forensic science. Specifically, the aim of this study was to implement advanced statistical methods to selected linguistic variables in forensic text comparison. In this respect, the methodology comprised a qualitative analysis and a quantitative analysis grounded on multivariate classical statistics, which can be defined as a simultaneous statistical analysis of a collection of variables and probabilistic methods such as the Likelihood-Ratio framework.

Corpus

One important concern was how to gather a corpus which would be comparable to corpora in the forensic world (typically characterized by a small number of authors, a small number of samples and short texts). The corpus used in this study was designed taking into consideration the importance of the availability of all the relevant sociolinguistic data about the individuals. Therefore, it was possible to avoid the effect of errors in independent variables. Finally, we were able to include texts by 47 informants. All of them were university students. Their native languages are Spanish and Catalan and they qualify as fully balanced bilingual speakers of both languages, since they have equivalent knowledge of both languages at levels corresponding to those of native speakers of each language (Baetens, 1989). All informants were between 18 and 25 years old and came from the Autonomous Community of Catalonia (Spain).

With the aim of gathering a corpus comparable to the forensic reality, participants were given the description of six different situations – one every week – and told to produce a Spanish written threatening message of approximately 600 words with a medium-high level of violence that could be understood as a verbal threat or as actual physiological violence against the recipient of their letter. This procedure resulted in a process of homogenisation of the corpus.

Thus, we compiled two different corpora: one for the BRK (Table 1) and another for the LR (Table 2). The corpus for the LR includes 22 men and 25 women and two samples per individual. The corpus to obtain likelihood ratios comprises 100% of women and 6 letters per each author since informants of this gender displayed the most cooperative attitude and showed willingness to participate in the process all the way through.

Gender	N individuals	N samples per individual	N total group samples	Mean Number of words	Std. Deviation Number of words	Std. Error Mean Number of words
Male	22	2	44	495.66	198.345	29.902
Female	25	2	50	577.70	184.680	17.451

Table 1. BRK corpus distribution.

Gender	N individuals	N samples per individual	N total group samples	Mean Number of words	Std. Deviation Number of words	Std. Error Mean Number of words
Female	18	6	108	568.48	189.67	18.856

Table 2. LR corpus distribution.

Variables

A linguistic variable is the representation of a linguistic feature that can be expressed in different ways with the same meaning. The linguistic variables in this study took the following fundamental characteristics into account: the variable ought to be highly frequent and stratified (Labov, 1972), show a high inter-individual variability and a low intra-individual variability, and also be relatively easy to extract and calculate (Nolan, 1983: 11), its variants should be interchangeable in some contexts (Tagliamonte, 2006: 73) and, finally, each variable ought to be as independent of other variables as possible (Rose, 2002: 52).

We also considered variables whose discriminatory potential had been evaluated in previous studies like Grant and Baker (2001), Chaski (2001), Wright (2013). And lastly, we considered variables which had been relevant in forensic linguistics casework carried out in the laboratory in which the author has worked.

A broad range of linguistic variables were analyzed and divided into four main groups: complexity, lexis, pragmatics and syntax. Table 3 shows a summary of the analyzed linguistic variables.

Complexity	Lexis	Pragmatics	Syntax
Number of words	Swearwords per sample	Intensification of the subject	Type of clause
Number of different words	Errors per sample	Expressing emphasis	Type of complex clause
Number of sentences	Expressing future	Number of questions	Type of juxtaposed clause
Number of paragraphs	Expressing obligation	Number of exclamations	Type of coordinated clause
Average sentence length [words]	Expressing condition	Addressing forms	
Average paragraph length [words]	Relative clauses	Greetings	
Average Word length [characters]	Subjunctive forms	Farewells	
TypeToken ratio (words)		Words in brackets	

Table 3. Summary of the analyzed variables.

Complexity measures analyzed in this study include the number of words per document, vocabulary richness (number of different words), the number of sentences and paragraphs, average lengths for sentences, paragraphs and words, and type-token ratio. This group was the only one analyzed semi-automatically by a perl code designed *ad hoc* and reviewed manually. The remaining groups were analyzed manually by the researcher.

In the analysis of lexis, frequencies of swearwords and errors per sample were calculated. Other features considered were whether the author used *ir a* + infinitive or the future tense to express future, *deber* + infinitive or *tener que* + infinitive to express obligation and whether the author used *como* or *si* to express condition.

Concerning the field of pragmatics, the distribution – presence or absence – of the first person singular personal pronoun, i.e. *yo*, was calculated in order to identify its intensification when present. The different ways of expressing emphasis such as capitalization, repetition or punctuation were also considered. Other pragmatic variables were the number of exclamations and interrogations used, the formality or informality of addressing pronouns, and the types of greetings and farewells, since they are reported by previous studies as possible authorship markers Wright (2013). Finally, the use of brackets to interject other text was evaluated.

Syntax was analyzed through an observation of the clause types used by the authors, i.e. complex or simple clauses, types of complex clauses – coordinated, juxtaposed or subordinated – and types of juxtaposed or coordinated clauses.

Method

This study proposes a combination of qualitative and quantitative approaches. Schmied (1993) notes that a "qualitative analysis is often a precursor for quantitative analysis, since before linguistic variables can be classified and counted, the categories for classification must first be identified".

During the qualitative analysis, linguistic features were identified in the data but no attempt was made to assign frequencies to those linguistic features. Instead, ambiguities inherent to the Spanish language were recognized. For instance, the word 'que' in Spanish (*that* in English) can be used in a corpus as a relative pronoun or as a conjunction. In contrast, features were classified and counted during the quantitative analysis. The measurement of the distribution of and the correlation between features led to the identification of characteristics which are likely to be genuine of the writer and therefore representative of his/her 'idiolectal style' and which reflect the author's behavior.

The statistical analysis was divided into two stages. The first stage consisted of the application of multivariate statistical techniques, which constitute an improvement of univariate analysis because "it incorporates information into the statistical analysis about the relationships between all the variables", according to Izenman (2008: 1).

But quantifying the degree of similarity is not enough for our purposes. As stated above, one must also consider the rarity or the expectancy of the distribution and correlation of features found to be similar between corpora in relation to the relevant population. This comparison can be addressed by the use of probabilistic methods such as the Likelihood-Ratio framework which carries out rigorous empirical analyses. Therefore, the second statistical stage consisted on the implementation of the LR framework.

Many researchers and practitioners state that the LR framewok is very well-suited to present evidence in court because it only weighs the impact of the evidence studied by the expert and it does not consider the court's prior or posterior beliefs. Aitken *et al.* (2011) state:

> To form an evaluative opinion from a set of observations, it is necessary for the forensic scientist to consider those observations in the light of propositions that represent the positions of the different participants in the legal process. The ratio of the probability of the observations given the prosecution proposition to the probability of the observations given the defence proposition, which is known as the likelihood ratio, provides the most appropriate foundation for assisting the court in establishing the weight that should be assigned to those observations. (p.1)

In this particular study, in order to obtain classification and subsequently the LR, we calculated the proximity distances among the author's samples (inter-variability) and also the distances within the author's samples (intra-variability). To calculate posterior probabilities for classification four algorithms of calculation were performed by discriminant analysis on the standard deviation of the distances with continuous variables.

The likelihood ratio was calculated considering four different classification tests:

- True positive: number of samples classified as belonging to their real author. 6 possible cases.
- False positive: number of samples classified as belonging to another author. 102 possible cases.
- True negative: Number of samples which are not classified as belonging to an incorrect author. 102 possible cases.
- False negative: Number of samples which are not classified as belonging to their real author. 6 possible cases.

Based on the results of these tests, the validity of the classifications was determined through the use of sensitivity and specificity tests. Sensitivity was defined as the probability of detecting an author's own samples and specificity as the probability of detecting samples that were not produced by that author, that is, the probability of rejecting foreign samples.

The subsequent step was to calculate the LR for each individual and for each of the variables in order to know the probability of the results. In particular, there were two ways to measure the likelihood ratio in this study, positively and negatively:

- Positive likelihood ratio (LR +) is the ratio between sensitivity and difference, that is, the probability that a sample is assigned to its author compared to the probability of a sample not produced by that author also being assigned to him or her.
- Negative likelihood ratio (LR–) is the ratio of the difference and specificity, that is, the probability that a sample is not assigned to its author compared to the probability that the rest of the samples are assigned to the rest of the authors.

$$LR+ = \frac{Sensitivity}{1\text{- }Specificity} = \frac{\frac{TP}{(TP + FN)}}{1 - \left(\frac{TN}{(TN + FP)}\right)}$$

$$LR- = \frac{1\text{- }Sensitivity}{Specificity} = \frac{1 - \left(\frac{TP}{(TP + FN)}\right)}{\left(\frac{TN}{(TN + FP)}\right)}$$

Figure 1. Formulas of the Likelihood ratio.

LR+ varies between zero and infinity – the higher its value, the greater the probability of classifying the unknown sample correctly. LR– varies between 0 and 1 – the lower the value, the greater the probability of correctly classifying the unknown sample. In order to assign the unknown sample to its author, these two conditions had to be fulfilled: an LR+ as high as possible and a LR– as low as possible. Thus, an author's samples were classified correctly when they met the following requirements: the group of samples that are classified correctly to their group (true positives) is large; the value of LR+ is very high (> 1000) and the value of LR– is minimal (0).

Summing up, qualitative analysis provides greater richness and precision, whereas quantitative analysis provides statistically reliable and generalizable results (McEnery and Wilson, 2001: 77).

Results

Base Rate Knowledge results

For each of the variables, a population distribution was provided, that is, the most commonly used variant of each variable and the expected frequency of that variant were established. A frequency rate higher or lower than that established by the population distribution may signal a particular characteristic of that author.

For instance, it was observed that the threat letters in the study were not abundant in abbreviations. Nevertheless, the distribution of the abbreviation of 'euros' was considered relevant because of its frequency in extortion letters from real cases. Results showed that the most common way of writing 'euros' in this corpus is in its non-shortened form (64.56%), followed by the sign '€' (33.33%) and, finally, the abbreviation 'EUR' (2.08%).

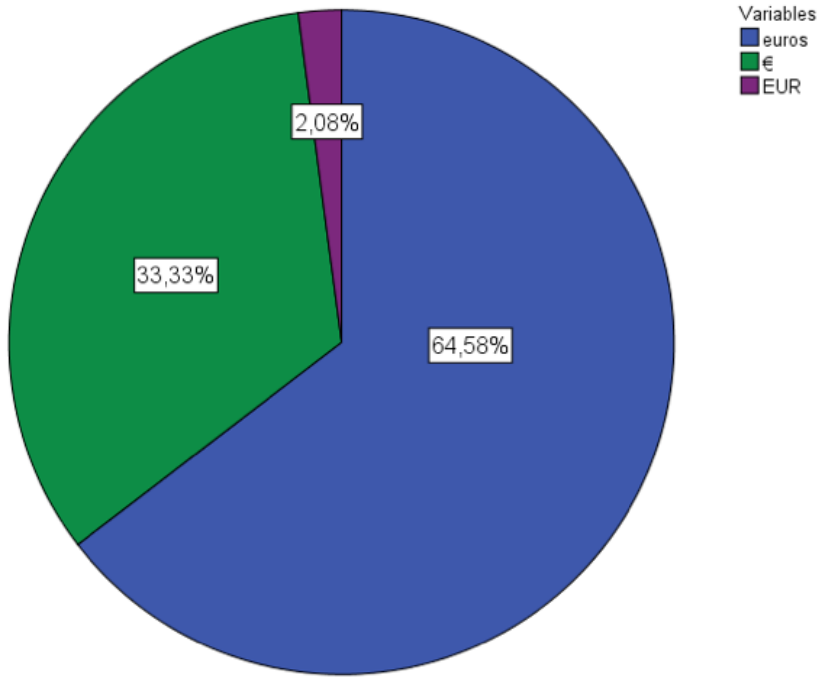

Figure 2. BRK of the substantive euros.

The way a speaker expresses emphasis may also differ in a relevant manner (Figure 3). In this study we analyzed the expression of emphasis by capitalization, the use of punctuation marks and the use of repetition. Results showed that the most common way of expressing emphasis in written texts is capitalization (70.48%), followed by punctuation marks (19.05%) and, finally, by using the repetition of words or expressions (10.48%).

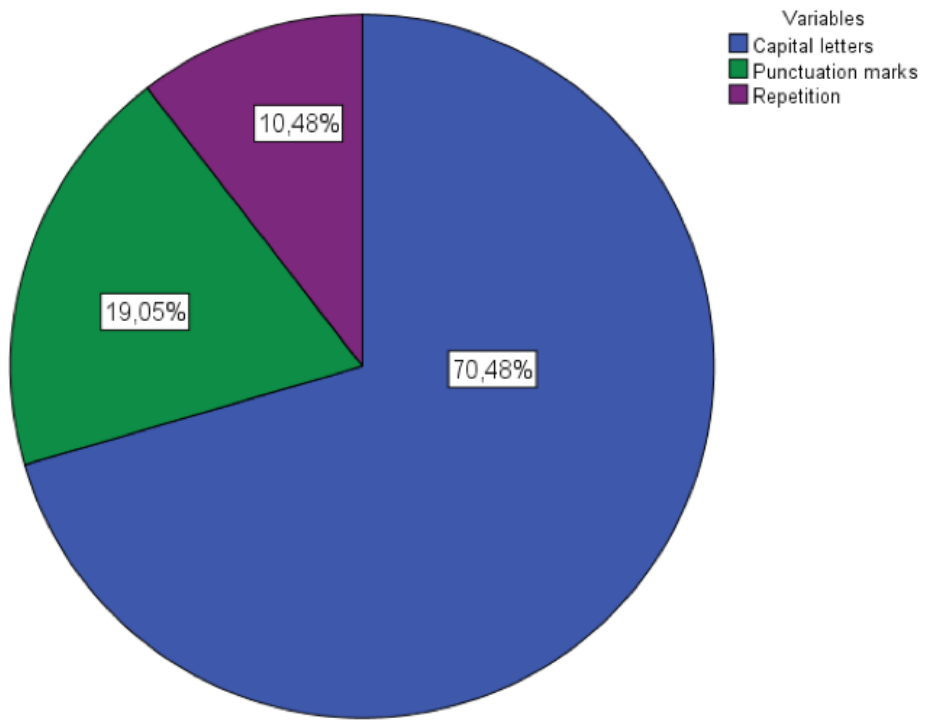

Figure 3. BRK expression of emphasis.

Population distributions of linguistic variables are useful for authorship attribution since they allow the expert to know the mean frequency values for each variant of the variables and, therefore, to what degree a variant may be expected to occur generally. Figure 4 shows the variable of lexical errors and the values for each individual sample. It is worth noting the individual behavior of certain writers. For example, writer 44 often makes significantly more errors than the average population, which is why, in the case of spelling errors, diacritics, and grammatical pleonasms, this author's samples are placed in the extreme values of the graph. Other cases of special interest are those in which the writer often makes a greater number of errors of a single type. For example, writer 35 shows a remarkable number of errors caused by the contact between Spanish and Catalan languages in both samples, and writer 41 shows some difficulties with normative punctuation. Extreme values are indicated with an asterisk and outliers with a circle.

Another example of BRK results is the variable of expression of obligation in Spanish shown in Figure 5. It is relevant to note the cases of authors 32 and 28. Author 32 stands out for using *deber* + infinitive frequently in both samples, while author 28 is the only author who uses *haber de* + infinitive.

Variables with discriminatory potential

Once the Base Rate Knowledge was established, the variables that offered a greater discriminatory potential were selected. Those variables showed low intra-individual variation and high inter-individual variation, thus, it should be possible to distinguish samples among individuals. Table 4 comprises the most discriminating variables.

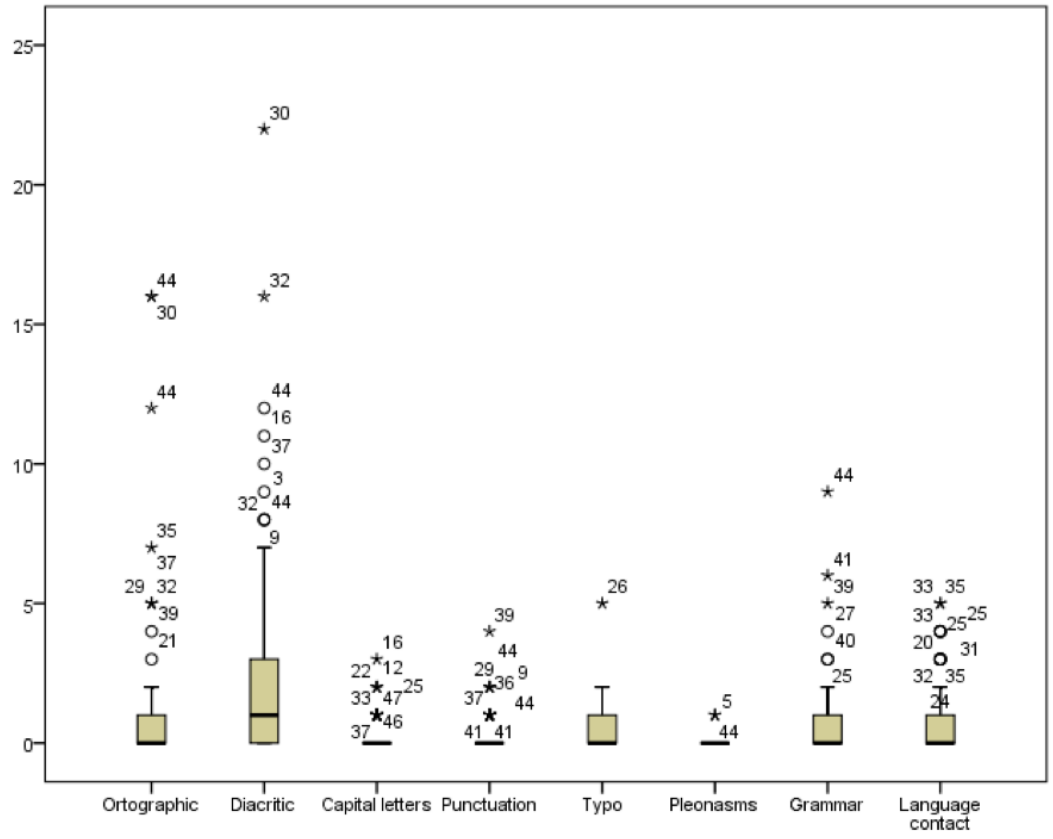

Figure 4. BRK lexical errors.

Complexity	Lexical	Syntax	Pragmatic
Number of paragraphs	Number of orthographic errors	Number of subordinate sentences	"Sincerely" (*Cordialmente*, ES)
Number of different tokens	Number of swearwords	Number of simple sentences	Absence of singular first person subject
Number of sentences	Number of language contact errors	Number of complex sentences	
Words per paragrah			

Table 4. Variables with discriminatory potential.

This set of variables was used to calculate the probabilities of success and failure in posterior classifications.

Likelihood Ratio results

Table 5 shows the classification results. Cells shaded in red show four authors who are completely different from the rest because all their samples (6) are classified correctly (meaning true positive) and no samples are attributed to another author (represented in

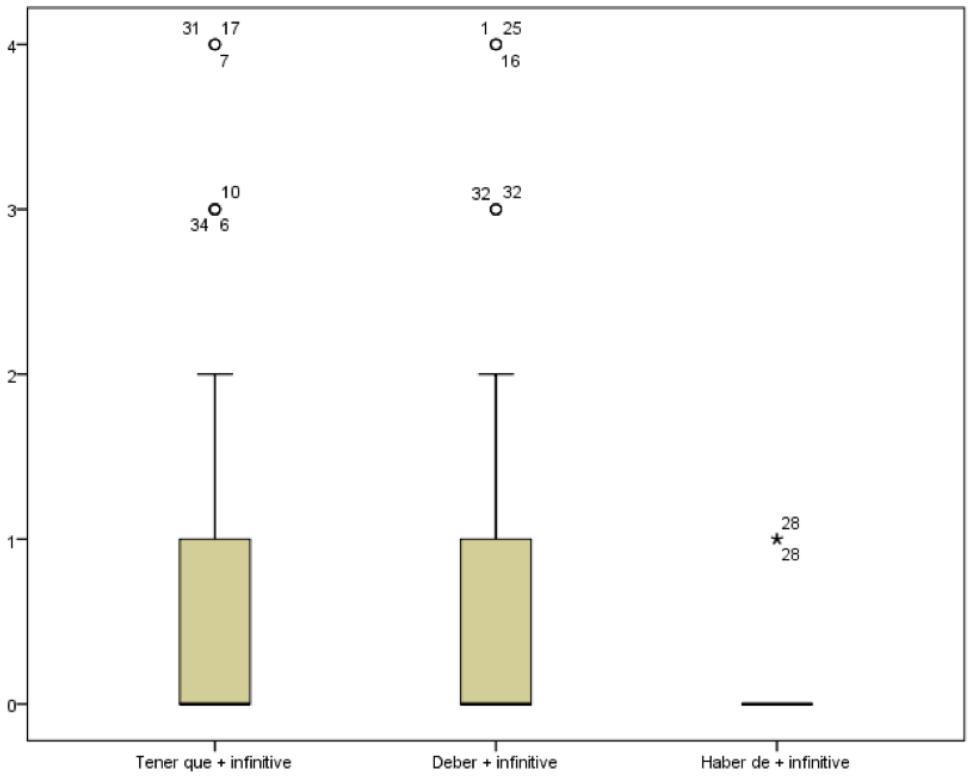

Figure 5. BRK expression of obligation.

the table as false positive). Orange cells indicate 6 authors who are well differentiated from the rest but share more features with other authors and, therefore, some of their samples are attributed to other authors.

Author	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
True positive	6	1	4	6	1	6	6	4	6	3	5	3	6	2	6	6	6	6
False positive	3	2	2	0	0	2	0	7	0	3	9	8	3	2	6	6	4	0
False negative	0	5	2	0	5	0	0	2	0	3	1	3	0	4	0	0	0	0
True negative	99	100	100	102	102	100	102	95	102	99	93	94	99	100	96	96	98	102

Table 5. Classification results.

The validity of the classifications must be determined from these figures, that is, to what extent the classifications obtained would fit more complex and rigorous processes.

Each position on the X axis of Figure 6 represents an individual and on the Y axis the probability of each of the samples. In green we can observe samples which are classified correctly to their author and in red samples which are not classified to the correct author (the number indicates the author which is incorrectly classified), that is, the method's sensitivity. Thus, this graph visually summarizes the probability of detecting an author's own samples.

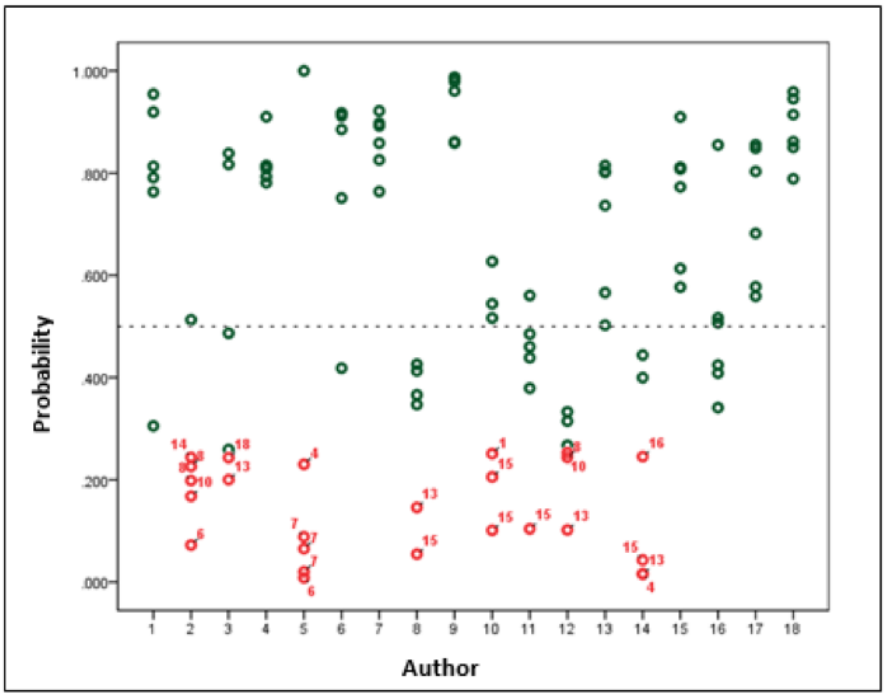

Figure 6. Sensitivity of the method.

According to these results, the classification potential is up to 76.85% and in more than half of the cases the classification probability is greater than 50%. It is important to highlight that all false negatives are below 25% probability and that all true positives are above this probability value.

With regard to the method's specificity, Figure 7 shows samples which are correctly classified in green and samples which are classified as belonging to an incorrect author in orange (notice that the number indicates the real author of the sample).

In this case, true positives are also situated in the higher odds. Furthermore, most of the true positives (77.94%) are above 50% probability and most of the false positives (82.45%) are below that percentage. However, specificity results are not as satisfactory as sensitivity results because false positives are above 25% and even 50%.

Table 6 shows the likelihood ratio results: 5 authors with a maximum positive LR (this value is denoted as > 1000), 10 with minimal negative LR (0.00) and 4 authors with a maximum and a minimum LR+ and LR– respectively.

Thus, the results so far complement recent advances in authorship attribution using LR with the integration of BRK. For example, Ishihara (2017) used word- and character-based features to attribute chatlog messages of different length by 115 authors and estimated the strength of this attributions with LR. The results of his model show a discrimination accuracy of around 76% with the shortest texts (500 words) and of around 94% with the longest (2500 words). On a different study, Ishihara (2014) applied an N-gram language model to a corpus of text messages, again divided into four groups of different sizes.

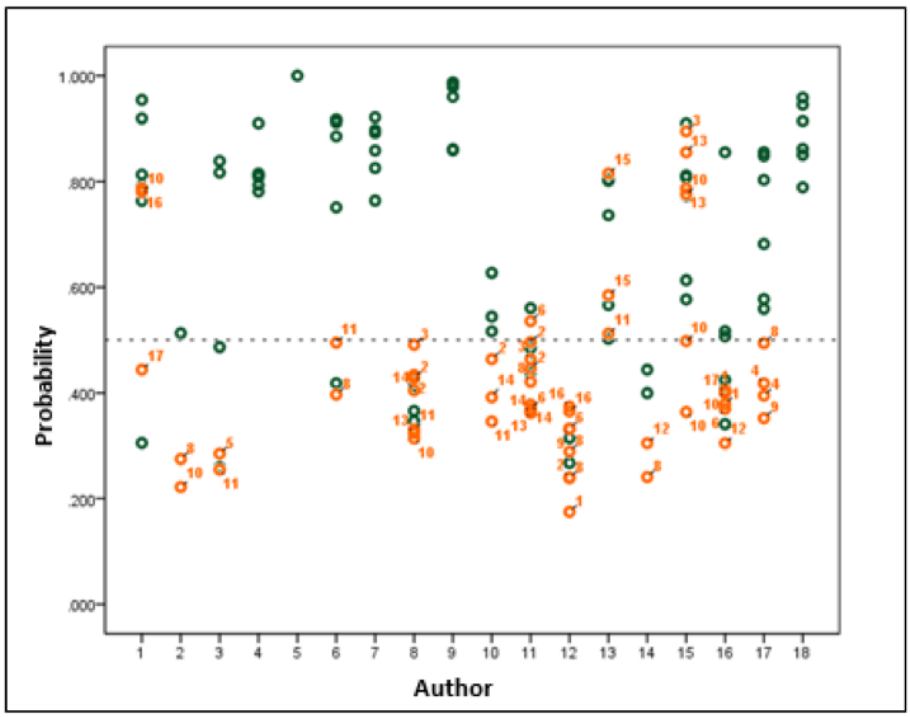

Figure 7. Specificity of the method.

Author	1	2	3	4	5	6	7	8	9
LR+	34.00	8.50	34.00	>1000	>1000	51.00	>1000	9.71	>1000
LR-	0.00	0.85	0.34	0.00	0.83	0.00	0.00	0.36	0.00
Author	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
LR+	17.00	9.44	6.38	34.00	17.00	17.00	17.00	25.50	>1000
LR-	0.52	0.18	0.54	0.00	0.68	0.00	0.00	0.00	0.00

Table 6. LR results.

Conclusions

This technique has correctly classified 75% of the samples, 60% of which with a probability greater than 50%. Finally, it should also be noted that there is a 25% sensitivity threshold since all the texts classified as belonging to their true author are above the 25% threshold and all the texts incorrectly classified are placed below this value.

As mentioned in the Introduction, the Committee on Identifying the Needs of the Forensic Sciences Community at the National Research Council of the United States published a document titled 'Strengthening Forensic Science in the United States: A Path Forward' (2009) which states:

> For decades, forensic sciences have produced valuable evidence that has contributed to the successful prosecution and conviction of criminals as well as to the examination of innocent people. Over the last two decades, advances in some forensic science disciplines, especially the use of DNA technology, have

> demonstrated that some areas of forensic science have great additional potential to help law enforcement identify criminals. Many crimes that may have gone unsolved are now being solved because forensic science is helping to identify the perpetrators. (p.26)

This statement must make the forensic scientific community realize its important role in society and therefore the – positive and negative – implications of its expert evidence. Due to the importance of the forensic expert's task, the community ought to set up a reliable methodology with agreed-upon standards and "should establish a professional body that not only promotes these goals but also certifies experts and, where applicable, accredits training programs and laboratories" (Koehler, 2013: 537).

At a general level, this study can contribute to forensic linguistics and particularly to the field of forensic text comparison, since the proposed methodology can be useful when resolving cases of authorship attribution and the corpus, the variables selected and the methodology may also represent a contribution to Corpus Linguistics, Computational Linguistics and the Likelihood-Ratio framework. Admittedly, however, this corpus has a relatively small number of participants to represent a comparative baseline to establish similar BRK and LR values for another language, which constitutes a significant limitation of the study. Additionally, as is commonly the case with research in forensic linguistics, any conclusions drawn from this study must consider the fact that the samples analyzed were produced in artificial contexts and that texts produced naturally would provide possibly provide more realistic information as to the authors' styles.

At a more detailed level, the most important contributions of this proposal have to do with the compilation of unified database of real-world texts in Peninsular Spanish in order to achieve a population distribution of linguistic variables in threatening letters; a common statistical method based on advanced multivariate statistical methods and the LR framework; a further small step towards the establishment of a code of good practice in forensic text comparison since control factors are considered during the collection of data, there are sampling procedures and qualitative and quantitative methods implemented. The implementation of a code of good practice can help to provide more reliable and conclusive results in authorship attribution.

Notwithstanding these results, there is still much to be done in the field of authorship attribution to reach the precision levels of the results of other forensic sciences taking into account the limits imposed by the nature of the object analyzed. It is necessary to develop and test new approaches to achieve comparable results taking into account the Achilles' heel of each research, for instance, the variability inherent in language.

References

Abercrombie, D. (1969). Voice qualities. In N. N. Markel, Ed., *Psycholinguistics: An introduction to the study of speech and personality*. London: The Dorsey Press.

Aitken, C., Berger, C. E., Buckleton, J. S., Champod, C., Curran, J., Dawid, A. and Jackson, G. (2011). Expressing evaluative opinions: A position statement. *Science and Justice*, 51(1), 1–2.

Aitken, C. G. G. and Taroni, F. (2004). *Statistics and the evaluation of evidence for forensic scientists*. Chichester, UK: John Wiley & Sons, john wiley ed.

Baetens, H. (1989). *Principis bàsic del bilingüisme*. Barcelona: La Magrana.

Bel, N., Queralt, S., Spassova, M. and Turell, M. T. (2012). The use of sequences of linguistic categories in forensic written text comparison revisited. In S. Tomblin, N. MacLeod, M. Coulthard and R. Sousa-Silva, Eds., *Proceedings of the International Association of Forensic Linguists' Tenth Biennial Conference*, 192–209, Birmingham, UK: Aston University Centre for Forensic Linguistics (Online).

Biber, D. (1988). *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D. (1995). *Dimensions of Register Variation: a Cross-Linguistic Comparison*. Cambridge; New York: Cambridge University Press.

Biber, D., Conrad, S. and Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Broeders, A. P. A. (1999). Some observations on the use of probability scales in forensic identification. *Forensic Linguistics*, 6(2), 228–241.

Burrows, J. (2003). Questions of authorship: Attribution and beyond A lecture delivered on the occasion of the roberto busa award ACH-ALLC 2001, new york. *Computers and the Humanities*, 37(1), 5–32.

Burrows, J. F. (1987). Word patterns and story shapes: The statistical analysis of narrative style. *Literary and Linguistic Computing*, 2, 61–70.

Chambers, J. K. (2009). *Sociolinguistic theory: linguistic variation and its social significance*. Oxford: Wiley-Blackwell.

Champod, C. and Evett, I. W. (2007). Commentary on APA Broeders (1999) 'Some observations on the use of probability scales in forensic identification'. Forensic Linguistics 6 (2): 228–41. *International Journal of Speech Language and the Law*, 7(2), 239–243.

Chaski, C. E. (2001). Empirical evaluations of language-based author identification techniques. *Forensic Linguistics*, 8(1), 1350–1771.

Cicres Bosch, J. (2007). *Aplicació de l'anàlisi de l'entonació i de l'alienació tonal a la identificació de parlants en fonètica forense*. Phd thesis, Universitat Pompeu Fabra, Spain.

Committee on Identifying the Needs of the Forensic Sciences Community - National Research Council, (2009). *Strengthening forensic science in the united States: A path forward*. Rapport interne.

Coulthard, M. (2004). Author Identification, Idiolect and Linguistic Uniqueness. *Applied Linguistics*, 25(4), 431–447.

Coulthard, M. and Johnson, A. (2007). *An Introduction to Forensic Linguistics: Language in Evidence*. London and New York: Routledge.

Evett, I. (1998). Towards a uniform framework for reporting opinions in forensic science casework. *Science and Justice*, 38(3), 198–202.

Fenton, N. and Neil, M. (2012). On limiting the use of bayes in presenting forensic evidence. 1–27.

Gavaldà Ferré, N. (2011). Sociolingüística de la variació i lingüística forense. *Llengua, Societat i Comunicació*, 9(49-59).

Gonzalez-Rodriguez, J., Drygajlo, A., Ramos-Castro, D., Garcia-Gomar, M. and Ortega-Garcia, J. (2006). Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition. *Computer Speech and Language*, 20(2–3), 331–355.

Grant, T. (2010). Txt 4n6: Idiolect free authorship analysis. In M. Coulthard and A. Johnson, Eds., *Routledge Handbook of Forensic Linguistics*. Routledge.

Grant, T. D. and Baker, K. L. (2001). Identifying reliable, valid markers of authorship: A response to chaski. *International Journal of Speech, Language and the Law*, 8(1), 66–79.

Guy, G. (1980). Variation in the group and the individual. In W. Labov, Ed., *Locating language in time and space*. New York: Academic Press, 1–36.

Holmes, D. I. (2003). Stylometry and the civil war: The case of the pickett letters. *Chance*, 16(2), 18–25.

Ishihara, S. (2014). A likelihood ratio-based evaluation of strength of authorship attribution evidence in SMS messages using N-grams. *International Journal of Speech, Language & the Law*, 21(1).

Ishihara, S. (2017). Strength of forensic text comparison evidence from stylometric features: a multivariate likelihood ratio-based analysis. *International Journal of Speech, Language & the Law*, 24(1).

Izenman, A. J. (2008). *Modern multivariate statistical techniques: Regression, classification, and manifold learning*. Berlin and Heidelberg: Springer.

Johnstone, B. (1996). *The linguistic individual: Self-expression in language and linguistics*. Oxford: Oxford University Press.

Juola, P. (2006). Authorship attribution. *Foundations and Trends in Information Retrieval*, 1(3), 233–334.

Koehler, J. J. (2013). Linguistic confusion in court: Evidence from the forensic sciences. *Brooklyn Law School's Journal of Law & Policy*, 21(2), 515–540.

Koppel, M., Schler, J. and Argamon, S. (2009). Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, 60(1), 9–26.

Labov, W. (1972). *Sociolinguistic patterns*. Oxford: Basil Blackwell.

McEnery, T. and Wilson, A. (2001). *Corpus linguistics*. Edinburgh: Edinburgh University Press.

Nolan, F. (1983). *The phonetic bases of speaker recognition*. Cambridge: Cambridge University Press.

Queralt, S. (2014). Acerca de la prueba lingüística en atribución de autoría hoy. *En Revista de Llengua i Dret*, 62, 35–48.

Queralt, S., Spassova, M. and Turell, M. T. (2011). L'ús de les combinacions de seqüències de categories gramaticals com a nova tècnica de comparació forense de textos escrits. *LSC- Llengua, societat i comunicació*, 9(59-67).

Queralt, S. and Turell, M. T. (2012). Testing the discriminatory potential of sequences of linguistic categories (n-grams) in Spanish, Catalan and English corpora. In *Regional Conference of the International Association of Forensic Linguists 2012*, Kuala Lumpur, Malaysia: University of Malaya.

Rose, P. (2002). *Forensic speaker identification*. London and New York: Taylor and Francis.

Schmied, J. (1993). Qualitative and quantitative research approaches to English relative constructions. In C. Souter and E. Atwell, Eds., *Corpus-based computational linguistics*. Amsterdam: Rodopi, 85–96.

Sjerps, M. and Biesheuvel, D. B. (2007). The interpretation of conventional and 'Bayesian' verbal scales for expressing expert opinion: A small experiment among jurists. *International Journal of Speech Language and the Law*, 6(2), 214–227.

Spassova, M. S. (2009). *El potencial discriminatorio de las secuencias de categorías gramaticales en la atribución forense de autoría de textos en español*. Phd, Universitat Pompeu Fabra, Spain.

Spassova, M. S. and Grant, T. D. (2008). Categorizing spanish written texts by author gender and origin by means of morpho-syntactic trigrams: Some observations on

method's feasibility of application for linguistic profiling. In *Curriculum, Language and the Law Inter-University Centre*, Dubrovnik: University of Zagreb.

Spassova, M. S. and Turell, M. T. (2007). The use of morpho-syntactically annotated tag sequences as forensic markers of authorship attribution. In M. T. Turell, M. S. Spassova and J. Cicres, Eds., *Proceedings of the second european IAFL conference on forensic linguistics, language and the law*, 229–237, Barcelona: Publicacions de l'IULA.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3), 538–556.

Tagliamonte, S. A. (2006). *Analysing sociolinguistic variation.* Cambridge: Cambridge University Press.

Turell, M. T. (1995). *La sociolingüística de la variació.* Barcelona: Promociones y Publicaciones Universitarias.

Turell, M. T. (2004a). Textual kidnapping revisited: The case of plagarism in literary translation. *International Journal of Speech Language and the Law*, 11(1), 1–26.

Turell, M. T. (2004b). The disputed authorship of electronic mail: Linguistic, stylistic and pragmatic markers in short texts. In *First European IAFL Conference on Forensic Linguistics, Language and Law*, Cardiff: Cardiff University.

Turell, M. T. (2010). The use of textual, grammatical and sociolinguistic evidence in forensic text comparison. *International Journal of Speech Language and the Law*, 17(2), 211–250.

Woolls, D. and Coulthard, M. (1998). Tools for the Trade. *International Journal of Speech, Language and the Law*, 5(1), 33–57.

Wright, D. (2013). Stylistic variation within genre conventions in the Enron email corpus: developing a textsensitive methodology for authorship research. *International Journal of Speech Language and the Law*, 20(1), 45–75.

The Rowling Protocol, Steven Bannon, and Rogue POTUS Staff: a Study in Computational Authorship Attribution

Patrick Juola

Juola and Associates, USA

Abstract. *A key step forward in the professionalization of forensic science is the development of standards of practice and protocols. Based on his analysis of the Rowling case, Juola (2015) proposed a systematic protocol for authorship verification. We present both a theoretical and an empirical analysis of the accuracy of this protocol. We further present a demonstration of this analysis in terms of a high-profile case of political activism. We show that this protocol produces accurate and understandable analyses of the likelihood of common authorship.*

Keywords: *Authorship attribution, standards, protocols, independence, statistical analysis.*

Resumo. *Um passo fundamental na profissionalização da ciência forense é o desenvolvimento de normas e protocolos de prática. Com base na sua análise do caso Rowling, Juola (2015) propôs um protocolo sistemático para verificação de autoria. Neste trabalho, apresentamos, quer uma análise teórica, quer uma análise empírica da precisão deste protocolo. Procedemos, ainda, a uma demonstração dessa análise em termos de um caso importante de ativismo político, mostrando que este protocolo permite produzir análises precisas e abrangentes da possibilidade de autoria comum.*

Palavras-chave: *Atribuição de autoria, normas, protocolos, independência, análise estatística.*

Introduction

The authorship of documents is a key question in many legal cases (both fictional and real), as a skim of many of Agatha Christie's mysteries will show.[1] Handwritten, or even typed, documents can be validated by physical marks of the production process.[2] Electronic documents (Chaski, 2005; Juola, 2006b, 2007) bring their own set of issues, as handwriting cannot be used to validate the documents, and one ASCII 'A' is bit-for-bit identical to any other. Stylometry, the study of individual writing style (Holmes, 1994; Grieve, 2005; Juola, 2006a; Stamatatos, 2009), can be so used. In the case of *Ceglia v. Zuckerberg, et al.* (McMenamin, 2011), for example, ownership of a significant part of Facebook depended in part on the validity of an emailed agreement between the two parties. By looking at the writing style, including aspects such as word choice, catch

phrases, punctuation and spelling, McMenamin was able to find the *linguistic* marks of the producer/author.

In this paper, we describe a specific type of authorship attribution problem, that of authorship verification, with some examples. We then describe a formal protocol based on the analytic techniques used to identify J.K. Rowling, the author of the *Harry Potter* novels, as the author of Robert Galbraith's *A Cuckoo's Calling* as well. We show how this protocol can be used to address a general and common class of problems and present a software system (Envelope) that implements this protocol in a simple and easy-to-use way. We present both a theoretical and an empirical analysis of the accuracy of this protocol. Finally, we present a demonstration of this analysis in terms of a high-profile case of political activism — that of "Rogue POTUS Staff," a political activist who ostensibly posts inside information about the Trump White House.

Background

Authorship Analysis

Language is among the most individualized activities people engage in. For this reason, much can be learned about a person by looking at his or her writings. An easy example is distinguishing between different regional groups. A Commonwealth English speaker/writer can easily be spotted by her use of "lorry" instead of "truck," spelling "labor" with a 'u,' and less obviously by grammatical constructions such as "could do" or "in hospital." These insights can be extended to questions of authorship without regard to handwriting. The basic theory of traditional stylistics is fairly simple. As McMenamin (2011) describes it,

> 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.

Coulthard's (2013) description is also apt:

> The underlying linguistic theory is that all speakers/writers of a given language have their own personal form of that language, technically labeled an idiolect. A speaker/writer's idiolect will manifest itself in distinctive and cumulatively unique rule-governed choices for encoding meaning linguistically in the written and spoken communications they produce. For example, in the case of vocabulary, every speaker/writer has a very large learned and stored set of words built up over many years. Such sets may differ slightly or considerably from the word sets that all other speakers/writers have similarly built up, in terms both of stored individual items in their passive vocabulary and, more importantly, in terms of their preferences for selecting and then combining these individual items in the production of texts.

These choices express themselves in a number of ways. In an expert witness report, McMenamin (2011) analyzed eleven different and distinct "features" of the writing in sets of both known (undisputed) and disputed emails. One feature, for example, hinged on the spelling of the word "cannot", and in particular whether it was written as one word

("cannot") or as two ("can not"). Another feature was the use of the single word "sorry" as a sentence opener (as opposed, for example, to "I'm sorry"). Coulthard (2013) similarly discussed (among other features) the use of the specific phrase "disgruntled employees." (Why "disgruntled" and not one of its myriad synonyms?) In both cases, significant differences in these features can be held as evidence of differences in authorship.

The legal implications of this type of evidence should be apparent. Chaski (2005) provides a dramatic example in a suspicious death. When a person's body is found next to a typed suicide note – in this case, it was typed into a computer, but it could just as easily have been typed on an actual typewriter – the specific machine used to produce the note is not in question. If the machine is shared (for example, used by several roommates in a house), fingerprint analysis may not reveal much except that the fingerprints of several people can be found on it. By analyzing the writing, Chaski was able to establish that the decedent was probably not the author of the suicide note, turning the apparent suicide into a murder, and enabling the police to eventually catch the perpetrator. But even without the drama, any case involving "anonymous" writing (such as poison-pen letters or emails) would be aided by the ability to find the actual author.

As typically defined (e.g. Mosteller and Wallace, 1964; Binongo, 2003), authorship "attribution" involves selecting the most likely author from a small but finite set of candidate authors. In the real world, cases often involve simply determining whether or not a single specific author wrote a single specific document, where the alternative answer is that the actual author is simply "someone else." Examples of this include Juola (2013c), Brooks and Flyn (2013) and Collins (2013). Authorship "verification," as this subproblem is called, is a more difficult task because there is no obvious way to assess the properties of millions or billions of potential authors who are not part of the document set. While a specific misspelling of "toutch" (Wellman, 1936) may be idiosyncratic to one person, that does not exclude the possibility of other, unrelated people also using that spelling.

Computational text analysis

Computer-based stylometry applies the same general theory, but with a few major differences. The basic assumption that people make individual choices about language still holds, but instead of ad hoc features selected by examination of the specific documents, the analysts use more general feature sets that apply across the spectrum of problems (Binongo, 2003; Burrows, 1989; Hoover, 2004; Koppel *et al.*, 2009; Juola, 2006a; Juola *et al.*, 2013; Mikros and Perifanos, 2013). Examples of feature sets include word use, character clusters, and so forth. Using these feature sets or others (Rudman, 1998), the features present in a document are automatically identified, gathered into collections of feature representations (such as vector spaces), and then classified using ordinary classification methods (Jockers and Witten, 2010; Juola, 2006a, 2012a; Koppel *et al.*, 2009; Noecker Jr. and Juola, 2009) to establish the most likely author.

Binongo (2003) provides a clear example of this. For background: the first fourteen books of the *Oz* series were written by L. Frank Baum before his death. After his death, the publisher approached another author, Ruth Plumly Thompson, to finish the then-incomplete (and arguably nonexistent) fifteenth book, *The Royal Book of Oz*. In his study of the authorship of the *Oz* books, Binongo collected the frequencies of the fifty most frequent words in English from the books of undisputed authorship (his feature set). He

applied principal component analysis (his classification method) to obtain a data visualization of the stylistic differences, then showed that the disputed 15th book clearly lay in a stylistic space corresponding to only one candidate author, Thompson. This would clearly be highly relevant evidence if the authorship (perhaps for copyright reasons) were being disputed in court.

From a legal standpoint, there are three key issues with this technology. The first, admissibility, has been addressed in detail elsewhere (Chaski, 2013; Coulthard, 2013; Juola, 2014, 2015) but is closely tied to the second issue, the scientific validity of the technology itself. Numerous surveys (Grieve, 2005; Jockers and Witten, 2010; Juola, 2006a; Koppel *et al.*, 2009; Stamatatos, 2009) and TREC-style conferences (Juola, 2004, 2012b; Juola and Stamatatos, 2013; Stamatatos *et al.*, 2014) have shown that authorship can be determined with high accuracy (typically 80% or better) using realistically-sized samples. Large-scale studies (Juola, 2012a; Vescovi, 2011) have confirmed that there are often many different "best practices" that perform well based on different features. This allows for ordinary data fusion techniques such as mixture-of-experts to boost accuracy rates to practical levels.

Types of Authorship Problem

The usefulness of the above technology has been demonstrated in actual disputes. Chase's murder case (Chaski, 2005) has already been mentioned. For example, Collins (2013) used a mixture of experts to validate a newly discovered short story by Edgar Allan Poe, and Juola (Brooks and Flyn, 2013; Brooks, 2013; Juola, 2013b) used a similar method to identify J.K. Rowling (the author of the *Harry Potter* series) as the author of the pseudonymously published detective novel *A Cuckoo's Calling*. In a legal context, Juola (2013c) was able to verify the authorship of anonymous newspaper columns in support of an asylum claim in a US immigration court. Finally, Grant (2013) was able to perform a similar analysis without the aid of computers and determine the identity of a murderer.

A detailed examination of these cases, however, reveals key differences among them. As typically defined (e.g. Mosteller and Wallace, 1964; Binongo, 2003; Grant, 2013), authorship "attribution" involves selecting the most likely author from a small but finite set of candidate authors. In the case studied by Grant, there were, realistically, only two actors of interest. This may be typical of crimes-of-person, where someone needs to be present to commit the crime, and only a small group of candidates (those, for example, who had physical access to the crime scene) need to be considered. In Grant's case, it is understood that without sophisticated technological spoofing, only a person with physical access to a cell phone can use that phone to send text messages. In this case, the task of the analyst is to assess the comparative similarity/likelihood of each possible candidate author for the documents in question. This so-called "closed class" task, then, does not need to consider "none of the above" as a serious contender, and is the simplest and easiest formulation of the problem of authorship attribution.

By contrast, cases often involve simply determining whether or not a single specific author wrote a single specific document, where the alternative answer is that the actual author is simply "someone else." This may be typical of the analysis of published documents, as the questioned manuscript might have been written from literally anywhere in the world. Similar issues arise with the analysis of electronically transmitted docu-

ments such as web pages and emails. Even an obvious idiosyncrasy may be shared with someone else thousands of kilometers away.

Work has been done in authorship verification such as the "imposter" method (Koppel and Winter, 2014), but their work is hard to use and to understand. First, their use of huge numbers (tens of thousands) of distractor authors may provide statistical power, but makes the task of data collection arduous and expensive. Second, the authors focus on one analysis repeated in a rather untransparent way, an analysis focusing on what a reviewer of this paper has correctly identified as perhaps the least understandable analysis method we ourselves use. Thirdly, their protocol relies on an undescribed ad-hoc cutoff threshold and does not lend itself well to intuitive odds judgements about what people are typically interested in, the actual likelihood that a given author wrote a specific work—the sort of intuitive presentation that is easily understandable to a judge or jury.

A Proposed Protocol

Juola (2014, 2015) presented a formal protocol for authorship verification and showed how it could be applied to several separate authorship disputes.

Key elements of this proposed protocol are:

- Suitable data for analysis, including an ad hoc set of distractor authors believed not to be connected to the case;
- A set of independent analysis methods that have been found to perform well on similar tasks;
- A predefined data fusion framework amenable to formal statistical analysis, so that the likelihood of error can be assessed mathematically;
- A predefined interpretation of the statistical results in human-understandable terms.

As an illustration, we here describe its application in the immigration case reported in Juola (2013c). The background is relatively straightforward; an immigrant, whose name and other identifying details have been changed for personal safety, was applying for asylum in the United States. Bilbo Baggins, as we have renamed him, was originally a citizen of Mordor, a successful journalist under his own name, but also an anonymous online critic of the Mordor government. He feared persecution for his political activities, but, of course, the political activities had not been performed openly under his own name. Could the author of these anonymous articles be linked with the articles published under Mr. Baggins' own name?

I was able to collect a set of 160 news articles by five different named authors, none of whom was Baggins. This, in turn, provided me with five separate "baseline document corpora" against which to compare the anonymous writings. Using the JGAAP software platform (Juola *et al.*, 2006, 2009),[3] stylistic "distances" (Noecker Jr. and Juola, 2009) were calculated between the anonymous documents and each of the candidate authors as well as Baggins' undisputed writings. These distances had been shown in prior work to be able to select (with relatively high accuracy) the correct author out of a set of candidate authors based on the principle that the smallest distance represents the most similar and therefore most likely author.

Should Baggins be the actual author of the anonymous articles, then, one would expect Baggins to be the closest author by distance measurement. In the event that, by

chance, I had selected the actual author of the anonymous articles as one of the distractors, we would expect Baggins not to be the closest, but that person instead. Should the actual author be a seventh person, not in the set (which is more likely than accidentally finding the actual author as a distractor), one would have no reason *a priori* to believe that Baggins is particularly likely to write using a similar style, so there is roughly one chance in six that he would be the most similar author.

Table 1. Potential outcomes of Baggins article analysis

Case 1		Case 2	
Position	Author	Position	Author
1	*Baggins*	1	Distractor 1
2	Distractor 1	2	Distractor 2
3	Distractor 2	3	Distractor 3
4	Distractor 3	4	*Baggins*
5	Distractor 4	5	Distractor 4
6	Distractor 5	6	Distractor 5
Probably Baggins		Probably not Baggins	

One can therefore describe the potential outcomes of this analysis in table 1. Case 1 describes a situation where Baggins is chosen as the closest and most likely author; in the event that Baggins is, in fact, the actual author, we would consider this the most probable case. Case 2 describes a situation where Baggins is not observed to be the closest author, which we would consider to be the most probable case in the event either that the true author had inadvertently been among the distractors (a highly unlikely coincidence) or that the actual author was not in our data set of known authors.

Thus, with high probability, we expect case 2 if Baggins is not the actual author, and case 1 only if either Baggins is the true author, or the unlikely event that the true author is someone who writes with a similar style to Baggins. If Baggins is not in the data set, then we would expect, by chance, case 2 to arise roughly $\frac{5}{6}$ of the time. Thus, if one treats "none-of-the-above" as the null hypothesis, we would have an effective p-value (for rejecting the null hypothesis) of 0.167 in case 1.

If greater confidence is desired, one can, however, improve upon these results using ensemble methods. Juola (2013c) wrote:

> The basic idea is the one behind getting a second opinion: if two (or more) independent experts agree in their analysis, our confidence in that result is increased (Juola, 2008). This can be formalized using probability theory: if the chance of an expert being right is x, the chance of her being wrong is therefore $(1 - x)$. (The chance of two such experts independently being wrong is $(1 - x)(1 - x)$ or $(1 - x)^2$, and in general, the chance of k experts all being wrong is $(1 - x)^k$. For example, if experts in general are right 90% of the time, the chance of one expert being wrong is 0.1 or 10%. The chance of two both being wrong is 0.01 or 1%, and for three experts, 0.001 or 0.1%. In [the Baggins analysis], the chance of our analysis being wrong, from above, is 16.7%. If a similar analysis yields the same result, the chance of them both being wrong is a mere 0.167 times 0.167, one chance in thirty-six, or about 2.78.

Repeating this test (as Juola did) with a second, independent analysis (and getting the same result) would give an effective p-value of roughly 0.0278, enough to reject the null hypothesis on a standard one-tailed cutoff of 0.05. Similarly, repeating this test a third time (as Juola did not), and again getting the same result would get an effective p-value of 0.00463. In rejecting the null hypothesis, he would thus have demonstrated evidence tending to show that it is highly unlikely that anyone other than Baggins wrote the disputed documents, and hence that Baggins is the true author of the questioned documents. This was, in fact, the outcome of the case, and Bilbo Baggins was permitted to remain in the United States.

Of course, there is no reason to restrict oneself to only two tests, and similarly no reason to restrict oneself to exactly five distractor authors. Similarly, a simple "first/not-first" cutoff may be impractical, but this test lends itself well to statistical tests such as Fisher's exact test (Fisher, 1971) applied to computed scores such as the rank sum of Baggins' positions. (The case above, for example, would be equivalent – under the null hypothesis – of rolling two dice to determine Baggins' score; the reader can confirm for himself that there is one chance in 36 of getting a rank sum of 2, and three chances in 36, less than one in ten, of getting a rank sum of 3 or smaller.)

As discussed in the following section, we have extended this proposed protocol to permit more accurate probability assessments by using more tests and a larger number of distractor authors. We have implemented this protocol in a software-as-a-service (SaaS) platform, named Envelope (Juola, 2016) to provide low-cost, high-accuracy resolution of authorship disputes.

Envelope, a SaaS Platform for Authorship Verification

Design and implementation

Envelope, in its current version, focuses on a specific (and relatively common) type of disputed document, electronic mail (Chaski, 2005; Coulthard, 2013; McMenamin, 2011) written in English. The system is presented with client-supplied copies of the disputed email(s) as well as samples known to have been written by the purported author. These documents are compared against a set of distractor authors (currently a set of ten gender-balanced authors extracted from the Enron corpus (Klimt and Yang, 2004)) and rank-ordered for similarity along five human-understandable features that have been shown to work well in large-scale testing (Juola, 2012a; Vescovi, 2011). The five measured dimensions are as follows:

- Authorial Vocabulary (Vocabulary overlap): Words are, of course, what a work is fundamentally all about. A crime novel is usually about a dead body and how people deal with the problem it poses, while a romance novel is about a small group of people and their feelings for each other. Even emails differ in word choices as discussed above (Coulthard, 2013; McMenamin, 2011; Juola, 2013a). Authorial vocabulary is also one of the best ways to tell individual writers apart, by looking at the choices they make, not only in the concepts they try to express, but the specific words they use to create their own individual expression. The degree of shared vocabulary is thus a key authorial indicator. This was calculated using a modified Jaccard distance that does not take into account frequency distribution, and hence is sensitive only to the question of whether the author does or does not use a particular word token.

- Expressive Complexity (Word length): One key attribute of authors is, on the one hand, their complexity, and on the other, their readability. A precise author who uses the exact specific word to every event – "that's not a car, that's a Cadillac; that's not a cat, but a tabby" – will more or less be forced to use rarer words. These rarer words, by their very nature, are typically longer (Zipf, 1949). A large and complex vocabulary will naturally be reflected in longer words, producing a very distinctive style of writing. By tracking the distribution of word lengths (n.b.: the percentage of words with various lengths, not just the average word length, which is known not to perform well), we can assess the expressive complexity of a given author.
- Character n-grams: In addition to comparing words directly, scholarship has shown (Cavnar and Trenkle, 1994; Mikros and Perifanos, 2013; Noecker Jr. and Juola, 2009; Stamatatos, 2013) that comparison of the frequency spectra of character clusters (for example, four adjacent letters, whether as part of a word like "eXAMPle" or across two words as in "iN␣THe") is a useful way to assess document similarity. This allows matching of similar but not identical words, such as different forms of the same stem or words with similar affixes, and even preferred combinations of words. We used normalized cosine distance (Noecker Jr. and Juola, 2009) to compare the frequencies of various character n-grams.
- Function words: One of the most telling and oft-studied aspects of an individual writer is their use of function words (Binongo, 2003; Burrows, 1989; Hoover, 2004), the simple, short, common, and almost meaningless words that form a substantial fraction of English writing. These words thus provide a good indication of the tone of the writing and the specific types of relationship expressed throughout the manuscript. We evaluated function words by restricting our attention to the fifty most frequent words using normalized cosine distance as above.
- Punctuation: Although not necessarily linguistically interesting, and often the choice of editor instead of author, punctuation offers an insight into social conventions that have little effect on the meaning of the text. Because they have little effect, they are often freely variable between authors. For example, an author's decision to use an Oxford comma, their choice of marking extraneous material (for example, with commas, parentheses, or brackets), the way they split sentences with semicolons, periods, or comma splices, and even whether punctuation is put inside or outside quotation marks, do not change the meaning. In unedited documents (such as email), they therefore provide a strongly topic-independent cue to authorship that is not directly related to the other dimensions. (See Grant, 2013; McMenamin, 2011 for some non-computational examples.)

Numerical analysis of the Envelope **protocol**

Along each document, the eleven possible authors (ten implausible distractor authors plus one plausible suspect) are ranked from #1 (most similar/likely) to #11. The rank sum of the purported author across all dimensions is calculated and used to fuse the different analyses. For example, if the purported author scored as the most similar author on all five dimensions (the most compelling possible result), the rank sum would be five. The system then uses Fisher's exact test (Fisher, 1971) to determine a likelihood that the specific experimental result could have been obtained by chance.

In more detail, we consider the null hypothesis that the disputed document was not written by the purported author, and that there is, in fact, no relationship between them. Under this assumption, the purported author would rank anywhere from #1 to #11 (with equal probability), averaging at the sixth slot. Consistently appearing closer than the sixth slot, then, is evidence of systematic similarity between the two authors across a variety of independent stylometric variables. An unrelated person is unlikely to show this kind of systematic similarity, and hence if the calculated rank sum is small enough, we can reject the null hypothesis at any specific alpha cutoff desired. The system as currently developed uses standard cutoffs: if the p-value is 0.05 or less, we consider this to be "strong indications of common authorship," while trend-level values (p-value of 0.10 or less) are "indications of common authorship." "Weak indications" occur at p-values of 0.20 or less. Inconclusive or outright contraindications are handled appropriately. Indications of different authorship are handled at the other tail of the distribution; for example "strong indications of different authorship" are defined as a p value of 0.95 or greater. (From a theoretical perspective, of course, we would expect two unrelated authors to produce a p-value, on average, of 0.50; we thus acknowledge that the category names are biased somewhat against dissimilar authorship.)

Investigating independence

One major issue is the unwarranted independence assumptions implicit in the fusion framework. Two analyses are "independent" if knowing the outcome of one analysis gives you no information that would let you predict the other analysis. A classic example of this would be two well-shuffled decks of cards; drawing an ace from one deck tells you little about what you would get drawing from the other. By contrast, two draws from the same deck are not independent; if you draw an ace for your first card, there are fewer aces left to be drawn, and the odds of drawing an ace are lowered slightly . Similarly, the odds of drawing a queen are raised slightly. "Card counters" use this lack of independence to estimate odds in a professional gambling context.

In a forensic linguistics context, if we determine that the questioned document is closer in terms of word length distribution to unrelated distractor author A than it is to author B, does this imply that the questioned document will also be more similar in terms of punctuation to A than to B?

If method 1 is right 90% percent of the time, and method 2 is right 90% of the time, that does not mean that both methods will be wrong only one time in 100. That calculation (perhaps obviously) only holds if method 1 and method 2 are independent. However, method 1 and method 2 might *never* both be wrong, or, more worrisomely, if method 2 is a simple replication of method 1, method 2 might be wrong every time method 1 is wrong, so both are wrong a full 10% of the time. To properly validate this system will require analysis of potential inter-analysis dependencies and updating the fusion appropriately.

We can see some preliminary data from Juola's Baggins case (Juola, 2013c). As discussed above, the data were analyzed twice using two methods (that differed primarily in feature weighting). If these two methods are, indeed, independent, the fact that a particular distractor author is first in one condition would not provide any information about that author's position in the other. More formally, we would expect zero correlation between the rank-orders of the two conditions.

Table 2. Actual outcomes of Baggins article analysis

Condition 1	Condition 2
Baggins	Baggins
Distractor 1	Distractor 2
Distractor 2	Distractor 4
Distractor 4	Distractor 5
Distractor 5	Distractor 1
Distractor 3	Distractor 3

A quick inspection shows that these two orderings do not appear independent; for example, distractor 3 is last in both conditions. This is precisely as unlikely as Baggins being first in both analyses. Expressed more formally, the calculated rank-order correlation is 0.4, yielding a p-value of roughly 0.50. Given the small sample size, this is not strong enough to reject the hypothesis of no correlation, but it is not sufficient either to make one feel comfortable claiming that independence is likely, or even plausible.

In this case, the assumptions behind Fisher's exact test do not hold and the numerical calculations described above are not necessarily accurate. Further work is obviously required to assess the relative independence of any proposed methods for an Envelope-like system. The joint accuracy can be determined using more sophisticated fusion methods, but these methods are typically not easily understandable and not something one would wish to bring into court to present to a judge or jury. Alternatively, one can perform experiments to determine empirically the accuracy of such a system under controlled conditions and present the results of those experiments as an estimate of the overall accuracy of the analysis. This method, discussed in the following sections, provides us with an easily understandable assessment of the accuracy and therefore the weight to be given to any particular piece of evidence.

Discounting for a moment the potential issue of independence assumptions, the system is designed to be capable of delivering a sophisticated stylometric analysis quickly, cheaply, and without human intervention (thereby minimizing analyst bias effects).

Accuracy and Validation

System accuracy

To enhance validity, the system as implemented performs a data validation process. Both the known and disputed documents need to be of sufficient length (currently defined as $=< 200$ words), and cannot include header information (which can be picked up, for example, by looking for *From:* lines). Furthermore, the documents must be in English (Cavnar and Trenkle, 1994) (which we currently approximate by confirming the existence of "the" in the files). Violations of these conditions are documented in the generated report but do not prevent analysis; more sophisticated (and expensive) human-based analysis may be necessary in these circumstances. For example, stylometric analysis technology is known to transfer well between languages (Hasanaj, 2013; Hasanaj *et al.*, 2014; Juola, 2009), but a new distractor corpus would be necessary.

Preliminary testing: English-language email

The accuracy of this system has been tested on a variety of other email samples drawn from 20 additional authors in the Enron (Klimt and Yang, 2004) corpus. Out of 375 tri-

als, 179 produced "strong" indications of authorship, and all 179 (100%) were correct. Similarly, "Weak" indications of authorship were correct in 21 of 23 cases (91%). Only 2 cases showed just "indications", and 1 of those (50%) was correct, while the remaining 43 inconclusive cases could not be validated, but showed significant numbers of both same (6) and different (37) author pairs. Thus, as expected, this method does not return an answer in all cases, but when an answer is returned, the accuracy is very high.

Additional testing: English-language blogs

For a more extensive evaluation, we turned to the Blog Authorship Corpus (Schler *et al.*, 2006)[4]. This corpus provides the collected posts of nearly 20,000 bloggers, containing nearly 700,000 posts and 140 million words. We extracted approximately 8,000 blogs containing 300 or more sentences. We first extracted a fixed set (as per *Envelope* design of ten designated distractor authors and collected the first 100 sentences as distractor samples. For same-author tests, we randomly selected 4,000 blogs. From these blogs, we collected the first hundred sentences as a sample KD, and the last hundred sentences as a sample QD (thus providing maximal opportunity for topic and stylistic drift).

For different-author tests, we selected 4,000 additional blogs and paired them randomly in a daisy-chain structure. From these blogs, we used the first hundred sentences as a known document and the last hundred sentences from a different blog as a questioned document. No blogger appeared in both the same-author and different author tests. Aside from the distractor authors, no passage was analyzed more than once across the entire experiment suite.

This procedure yielded 4000 independent tests of same-author accuracy and of different-author accuracy. Table 3 shows the results, using the previously defined Envelope categories.

Table 3. Results of blog analysis under same- and different-author conditions

Result	Same-author	Different-author	Odds Ratio
Strong same	2,948	748	3.941
Same	246	359	0.686
Weak same	195	396	0.492
Inconclusive	409	1,390	0.294
Weak different	54	234	0.231
Different	47	230	0.204
Strong different	91	663	0.137

The final column of table 3 shows the odds ratio – the number of same-author attributions in that category divided by the number of different-author attributions in that category.

Application: The Case of "Rogue POTUS Staff"

We present one such case study, that of "Rogue POTUS ("President Of The United States.") Staff" (henceforth RPS), an anonymous political commentator, ostensibly from within the Trump White House staff.

Case background

The Twitter microblogging platform provides an easy way to publish short texts (140 characters, recently raised to 280) to a wide audience. It has become one of the largest social media platforms, with more than 330 million monthly active users (https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/) and more than a billion visits per month. Among those users is "Rogue POTUS Staff" (@RoguePOTUSSTAFF), self-described[5] as "The unofficial resistance team inside the White House. We pull back the curtain to expose the real workings inside this disastrous, frightening Administration."

Since even before its inauguration, the Trump administration has been plagued by controversy; staff turnover in the White House has been "off the charts" [6]. While historians will dig at the inner workings of the Trump White House for decades, one clear factor will be the tension between two groups of Trump supporters that are largely at odds, the traditional Republican establishment (such as former DIA director Flynn) and the alt-right ideologues (such as Bannon). Bannon's firing on August 18, 2017 was a result of just such a power struggle.

Approximately a month after Bannon's dismissal, the British online newspaper METRO published an article [7] suggesting that RPS was, in fact, a Twitter account run and written by Bannon himself. The primary evidence cited in this article was a suspicious correspondence in timing; RPS' last tweet was on August 16, two days before Bannon's firing. (After this time, no more tweets were issued from this account until October 29, more than a month after the article in question, and the October tweet contains no actual information, simply a threat: "Did you let silence become a false friend of security Mr. President? Tick tock, tick tock.") This is in marked contrast to previous activity — for example, on July 21, RPS posted twenty separate messages. Did something happen to RPS in mid-August?

Others have disputed this account, claiming that RPS is simply a hoax.[8] This, then, can be viewed as a classic instance of anonymous political discourse such as the publication by "Publius" of the Federalist Papers. The question of whether RPS and Bannon are the same author is thus a typical authorship verification problem, very similar to the question of whether "Publius" is the same person as Alexander Hamilton (Mosteller and Wallace, 1964) or whether Galbraith and Rowling are the same, and to be handled in a very similar way.

Materials and Methods

The Envelope method described in the previous sections was applied to answer this question.

As is typical, it was first necessary to collect undisputed samples of both RPS' writing style and Bannon's. Unfortunately, collecting writing sample data on Twitter is problematic. Much of Twitter is simply "re-tweets" of other people's writings (presumably in other people's style), and "almost 50% of (Twitter) traffic is generated and propagated by a rapidly growing bot population" (Juola *et al.*, 2018) (which again would not reflect the ostensible author's style). Bannon's official Twitter feed, in particular, consists almost exclusively of procedurally generated tweets. However, Bannon has written a lot of articles for the Breitbart media platform which are published under his own name. Similarly, RPS has a web presence (http://potusstaff.com/) containing, among other things,

editorial articles similar to those on Breitbart. We collected two RPS articles, and articles by Bannon as well as nine other Breitbart authors as distractors.

All analyses were performed using the Envelope engine. As described above, five sub-analyses were performed and folded into the overall system. To recap, we analyzed the authorial vocabulary (vocabulary overlap), expressive complexity (word lengths), character 4-grams, function words (the 50 most frequent words) and punctuation.

Our first comparison was of one RPS document to another, to confirm that RPS was, in fact, a single-author project (their use of "we" notwithstanding). One RPS document was used as the "questioned" or "unknown" document, and compared to eleven other documents (ten non-Bannon distractors and the second RPS document). If RPS were, in fact, a unitary author, then we would expect a ranking near 1 reflecting the stylistic uniformity. If RPS were not a unitary author, then there is little reason to suppose that the two RPS documents would be similar, and the expected rank might be anywhere from 1 to 11, averaging a 6. Fisher's exact test can measure the likelihood of a chance similarity with precision. Similarly, we compared Bannon's Breitbart articles to each of the two RPS articles separately, resulting in two additional results. All results are presented in the following subsection.

Results

Comparing the two RPS articles against each other produced a measured (theoretical) p-value of 0.008. Table 3 shows that when blog posts of comparable p-value ($p < 0.05$) are analyzed, the results are 4:1 that they are by the same author. We therefore conclude that these two articles (and by extension, the RPS editorials on http://potusstaff.com) are by the same, single author.

Conversely, the two RPS articles compared to the known Bannon article produced p-values of 0.6343 and 0.9729, respectively. In other words, not only were the articles not particularly similar, they were in fact more dissimilar than they were similar (more than half of the distractor authors were more similar). The odds ratio from table 3 are roughly 3:1 and 7:1 (respectively) against the articles being by the same person.

This demonstrates that there are substantial and robust stylistic differences between Bannon's writing and that of the (unknown) RTS author, while the style of the RTS author is uniform enough to allow us to believe him/her to be a single person. This strongly suggests that Metro was wrong and that "Rogue POTUS Staff" was not, in fact, Stephen Bannon. Our results further suggest that we can have high confidence in this finding.

Discussion

Precision and Recall

As was seen in the large-scale tests, nearly three-fourths of the actual same-author cases were identified as "strong indicators of common authorship," a recall rate of nearly 75%. Less than 2.5% were categorized as "strong indicators of different authorship." Thus documents actually by the same author are highly likely to be identified as such. Similarly, documents classified as "strong indicators of common authorship" included 2,948 correct attributions out of 3,696 so classified, a precision of 80%.

At the same time, documents by different authors are substantially less skewed, with the paradoxical result that "weak indications" or even merely "indications" of similar au-

thorship are actually less likely to arise from same-author analyses than from different-author analysis.

The odds-ratio clearly shows that, as the strength of indication of similar authorship goes down, the probability of similar authorship also decreases. The directionality of this relationship is perfect, in keeping with previous research (e.g. DeCarlo, 2013).

At the same time, it is also clear that, in contrast to theoretical predictions, the distribution of Fisher scores and by extension p-values in the different-author case is not uniform. For example, only 5% of the scores are expected to return p-values of 0.05 or less, while 18.7% actually did. This indicates, as discussed in the following section, a need for greater independence among the individual tests.

Genre effects

A second concern relates to the relationship between the calibration studies and RPS analysis; the calibration studies were done with blogs, not editorial articles. This genre will probably not directly affect our conclusion about RPS' identity, but may affect our confidence in unknown ways. One factor that should not be a concern are issues of representativeness in the distractor set, as recent research has shown that this does not affect accuracy very much (DeCarlo, 2013).

Finally, our analysis hinges crucially on the non-mathematical assumption, first, that the articles published under Bannon's by-line in Breitbart are actually by him and not by a ghost-writer, and similarly that the articles on RPS' web site are by the same RPS writer who writes the tweets. While we have shown some stylistic similarities in the RPS articles, there is no practical way to actually validate physical authorship. Of course, analysts have similar issues with many other authors; few if any modern scholars have seen a physical manuscript of Einstein's 1905 relativity paper, so there is no way to disprove the idea that it was written by someone else. In the absence of evidence to the contrary, we make the assumption that the claims of authorship mean what they say.

Conclusions

Despite these concerns, we feel the Envelope system delivers a high-quality analysis quickly and at low cost. Being fully automatic, the analysis is reproducible and is not influenced by analyst bias in any specific case. The probability of error has been confirmed empirically to be low (as expressed in table 3, and is lowest precisely when the analysis yields the strongest results. It is easy to extend the current system to additional languages, additional document types, or even additional classification tasks such as author profiling (Argamon *et al.*, 2009).

Interpreting an Envelope report is fairly straightforward; in the event of a "same author" finding, it means that, at the minimum, the actual author of the questioned document shared five human-understandable characteristics of writing style with the person who wrote the known document of interest. If the author of the will was not the decedent, it was, at a minimum, someone who used the same characteristic vocabulary, syntax, her characteristic style of punctuation, and even the function words in the same way. The computer can characterize the likelihood of this kind of match occurring with a person off the street using the statistics described above. As the old joke has it, "if it looks like a duck, walks like a duck, and uses punctuation like a duck…."

There are a number of fairly obvious extensions and possible improvements. Extension to new genres and/or languages (Hasanaj, 2013; Hasanaj *et al.*, 2014; Juola, 2009)

can be as simple as the creation of a new set of distractor documents. It may be possible to improve the accuracy by the incorporation of other analysis methods and feature sets (for example, the distribution of part-of-speech tags), although high-level processing such as POS tagging may limit its use in other languages. We continue our preliminary testing and will be expanding our offerings in terms of genre.

So, who did write the will, or at least (in the modern Christie remake) the Web posting? Who wrote the email ostensibly dividing up ownership of the startup, or revealing confidential business information? Computational analysis, as typified by Envelope, may not be able to provide definitive answers, but the evidence it creates can provide valuable information to help guide investigations or suggest preliminary conclusions. This system provides low-cost advice without the time and cost of human analysis, while retaining high accuracy.

Notes

[1]For those not familiar with Christie's work, an appropriate start might be *Peril at End House*.

[2]The reader is invited to look at Dorothy L. Sayer's *Strong Poison*.

[3]The JGAAP program is freely available as an open-source program; we have used it as well for the present study.

[4]See also `http://u.cs.biu.ac.il/~koppel/BlogCorpus.htm`

[5]See https://twitter.com/roguepotusstaff?lang=en.

[6]https://www.newyorker.com/news/news-desk/a-year-into-the-trump-era-white-house-staff-turnover-is-off-the-charts

[7]http://metro.co.uk/2017/09/29/was-rogue-white-house-twitter-account-actually-steve-bannon-6965449/

[8]Cf. https://theoutline.com/post/2396/trump-resistance-phonies

References

Argamon, S., Koppel, M., Pennebaker, J. W. and Schler, J. (2009). Automatically profiling the author of an anonymous text. *Communications of the ACM*, 52(2), 119–123.

Binongo, J. N. G. (2003). Who wrote the 15th book of Oz? an application of multivariate analysis to authorship attribution. *Chance*, 16(2), 9–17.

Brooks, R. (2013). Whodunnit? JK Rowling's secret life as wizard crime writer revealed. *Sunday Times*, 14 July.

Brooks, R. and Flyn, C. (2013). JK Rowling: The cuckoo in crime novel nest. *Sunday Times*, 14 July.

Burrows, J. F. (1989). 'an ocean where each kind…' : Statistical analysis and some major determinants of literary style. *Computers and the Humanities*, 23(4-5), 309–21.

Cavnar, W. B. and Trenkle, J. M. (1994). N-gram-based text categorization. In *1994 Symposium on Document Analysis and Information Retrieval*, 161–176.

Chaski, C. (2013). Best practices and admissibility of forensics author identification. *Journal of Law and Policy*, XXI(2), 333–376.

Chaski, C. E. (2005). Who's at the keyboard: Authorship attribution in digital evidence invesigations. *International Journal of Digital Evidence*, 4(1), n/a. Electronic-only journal: http://www.ijde.org, accessed 5.31.2007.

Collins, P. (2013). Poe's debut, hidden in plain sight. *The New Yorker*, October.

Coulthard, M. (2013). On admissible linguistic evidence. *Journal of Law and Policy*, XXI(2), 441–466.

DeCarlo, E. (2013). Inferring authorship through Myers-Briggs Type Inventory. In *Proceedings of DHCS 2013*, Chicago.

Fisher, R. A. (1971). *The Design of Experiments*. New York: Macmillan, 9th ed.

Grant, T. (2013). Txt 4n6: Describing and measuring consistency and distinctiveness in the analysis of SMS text messages. *Journal of Law and Policy*, XXI(2), 467–494.

Grieve, J. W. (2005). Quantitative authorship attribution: A history and an evaluation of techniques. Master's thesis, Simon Fraser University. URI: http://hdl.handle.net/1892/2055, accessed 5.31.2007.

Hasanaj, B. (2013). Authorship attribution methods in Albanian. In *Duquesne University Graduate Student Research Symposium*.

Hasanaj, B., Purnell, E. and Juola, P. (2014). Cross-linguistic transference of authorship attribution. In *Proceedings of the International Quantitative Linguistic Conference (QUALICO)*.

Holmes, D. I. (1994). Authorship attribution. *Computers and the Humanities*, 28(2), 87–106.

Hoover, D. L. (2004). Delta prime? *Literary and Linguistic Computing*, 19(4), 477–495.

Jockers, M. L. and Witten, D. (2010). A comparative study of machine learning methods for authorship attribution. *LLC*, 25(2), 215–23.

Juola, P. (2004). Ad-hoc authorship attribution competition. In *Proc. 2004 Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities (ALLC/ACH 2004)*, Göteborg, Sweden.

Juola, P. (2006a). Authorship attribution. *Foundations and Trends in Information Retrieval*, 1(3).

Juola, P. (2006b). Authorship attribution for electronic documents. In M. Olivier and S. Shenoi, Eds., *Advances in Digital Forensics II*, volume 222 of *International Federal for Information Processing*. Boston: Springer, 119–130.

Juola, P. (2007). Future trends in authorship attribution. In P. Craiger and S. Shenoi, Eds., *Advances in Digital Forensics III*, International Federal for Information Processing. Boston: Springer, 119–132.

Juola, P. (2008). Authorship attribution : What mixture-of-experts says we don't yet know. In *Proceedings of American Association for Corpus Linguistics 2008*, Provo, UT USA.

Juola, P. (2009). Cross-linguistic transference of authorship attribution, or why english-only prototypes are acceptable. In *Proceedings of Digital Humanities 2009*, College Park, MD.

Juola, P. (2012a). Large-scale experiments in authorship attribution. *English Studies*, 93(3), 275–283.

Juola, P. (2012b). An overview of the traditional authorship attribution subtask. In *Proceedings of PAN/CLEF 2012*, Rome, Italy.

Juola, P. (2013a). A critical examination of the Ceglia/Zuckerberg email authorship study. In *Proceedings of the 11th Biennial Conference on Forensic Linguistics/Language and Law of the International Association of Forensic Linguists (IAFL 2013)*, Mexico City, MX.

Juola, P. (2013b). How a computer program helped reveal J. K. Rowling as author of A Cuckoo's Calling. *Scientific American*, August.

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

Juola, P. (2014). The Rowling case: A proposed standard protocol for authorship attribution. In *Proceedings of Digital Humanities 2014*, Lausanne, Switzerland.

Juola, P. (2015). The Rowling case: A proposed standard protocol for authorship attribution. *DSH (Digital Scholarship in the Humanities)*.

Juola, P. (2016). Did Aunt Prunella really write that will? a simple and understandable computational assessment of authorial likelihood. In *Proc. A Workshop on Legal Text, Document, and Corpus Analytics (LTDCA 2016)*, 37–41.

Juola, P., Mikros, G. K. and Vinsick, S. (2018). Correlations and potential cross-linguistic indicators of writing style. *Journal of Quantitative Linguistics*, 26(2), 146–171.

Juola, P., Noecker, Jr. J., Ryan, M. and Speer, S. (2009). Jgaap 4.0 — a revised authorship attribution tool. In *Proceedings of Digital Humanities 2009*, College Park, MD.

Juola, P., Noecker Jr, J. I., Stolerman, A., Ryan, M. V., Brennan, P. and Greenstadt, R. (2013). Keyboard behavior-based authentication for security. *IT Professional*, 15, 8–11.

Juola, P., Sofko, J. and Brennan, P. (2006). A prototype for authorship attribution studies. *Literary and Linguistic Computing*, 21(2), 169–178. Advance Access published on April 12, 2006; doi: doi:10.1093/llc/fql019.

Juola, P. and Stamatatos, E. (2013). Overview of the authorship identification task. In *Proceedings of PAN/CLEF 2013*, Valencia, Spain.

Klimt, B. and Yang, Y. (2004). The Enron corpus: A new dataset for email classification research. *Machine Learning: ECML 2004*, 217–226.

Koppel, M., Schler, J. and Argamon, S. (2009). Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, 60(1), 9–26.

Koppel, M. and Winter, Y. (2014). Determining if two documents are written by the same author. *Journal of the Association for Information Science and Technology*, 65(1), 178–187.

McMenamin, G. (2011). Declaration of Gerald McMenamin. Available online at `http://www.scribd.com/doc/67951469/Expert-Report-Gerald-McMenamin`.

Mikros, G. K. and Perifanos, K. (2013). Authorship attribution in greek tweets using multilevel author's n-gram profiles. In *Papers from the 2013 AAAI Spring Symposium "Analyzing Microtext", 25-27 March 2013, Stanford, California*. Palo Alto, California: AAAI Press, 17–23.

Mosteller, F. and Wallace, D. L. (1964). *Inference and Disputed Authorship : The Federalist*, volume 58. Addison-Wesley.

Noecker Jr., J. and Juola, P. (2009). Cosine distance nearest-neighbor classification for authorship attribution. In *Proceedings of Digital Humanities 2009*, College Park, MD.

Rudman, J. (1998). The state of authorship attribution studies: Some problems and solutions. *Computers and the Humanities*, 31, 351–365.

Schler, J., Koppel, M., Argamon, S. and Pennebaker, J. (2006). Effects of age and gender on blogging. In *Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3), 538–56.

Stamatatos, E. (2013). On the robustness of authorship attribution based on character n-gram features. *Journal of Law and Policy*, XXI(2), 420–440.

Stamatatos, E., Stein, B., Daelemans, W., Juola, P., Barrón-Cedeño, A., Verhoeven, B. and Sanchez-Perez, M. A. (2014). Overview of the authorship identification task at PAN 2014. In *Proceedings of PAN/CLEF 2014*, Sheffield, UK.

Vescovi, D. M. (2011). Best practices in authorship attribution of English essays. Master's thesis, Duquesne University.

Wellman, F. L. (1936). *The Art of Cross-Examination.* New York: MacMillan, 4th ed.
Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort.* New York: Hafner Publishing Company. Reprinted 1965.

On the Implications of the General Data Protection Regulation on the Organisation of Evaluation Tasks

Francisco Rangel & Paolo Rosso

Autoritas Consulting, S.A., Spain & Universitat Politècnica de València, Spain

Abstract. *Evaluation campaigns allow for the creation of a common framework for research, making possible comparability and reproducibility in science. Furthermore, the huge amount of publicly available data in the different social platforms (social big data) favours evaluation tasks proliferation, for example in forensic linguistics. However, due to the implications that the release of the data may have on the privacy of people, rules for their protection must be laid down. These norms have been defined by the European Commission in the General Data Protection Regulation (GDPR) of April 27, 2016. Moreover, for the collection and distribution of data, each social media platform defines its legal base to use its data. In this paper, we describe the GDPR articles that apply for the organisation of evaluation tasks. Moreover, we propose a methodology to follow at the time of the organisation of evaluation tasks. Finally, we show a case study about the organisation of the PAN forensic linguistic tasks on author profiling at CLEF that we have been organising since 2013, showing how both GDPR and Twitter Terms of Service have been met when creating and distributing the corpora.*

Keywords: *GDPR, Corpora, Evaluation Tasks, Author Profiling.*

Resumo. *As tarefas de avaliação permitem a criação de um enquadramento de avaliação comum, permitindo a comparabilidade e reprodutibilidade na ciência. A enorme quantidade de dados disponíveis publicamente nas diferentes plataformas sociais (social big data) contribui para a proliferação das tarefas de avaliação, por exemplo na área da linguística forense. Contudo, decorrente das possíveis implicações da divulgação dos dados para a privacidade das pessoas, são necessárias regras para sua proteção. Estas normas foram definidas pela Comissão Europeia no Regulamento Geral de Proteção de Dados (RGPD) de 27 de abril de 2016. Além disso, para efeitos de recolha e distribuição dos dados, cada plataforma de rede social define a sua base jurídica para utilizar os seus dados. Neste artigo, descrevemos os artigos do RGPD aplicáveis à organização de tarefas de avaliação. Propomos, ainda, uma metodologia a seguir para organização de tarefas de avaliação. Finalmente, apresentamos um estudo de caso sobre a organização das tarefas de linguística forense do PAN no CLEF para determinar o perfil dos autores, que organizamos desde 2013, mostrando de que modo observamos, quer o RGPD,*

quer os Termos e Condições do Twitter, na criação e distribuição dos corpora.

Palavras-chave: *GDPR, Corpora, Tarefas de avaliação, Perfil dos autores.*

Introduction

It might be said that the main objective when organising evaluation tasks is to provide with a common framework where researchers can experiment and evaluate their results under the same conditions. Namely, a framework where both the data and the evaluation methodology are common to all the researchers. This evaluation framework allows for comparability and reproducibility.

The existence and publicly availability of big amounts of data in social platforms (namely social big data) favours the proliferation of evaluation tasks. This is also true in case of forensic linguistics Coulthard *et al.* (2016). In this vein, there are several evaluation tasks organised around the globe related to forensic linguistics. For example PAN,[1] the lab at CLEF[2] on digital text forensics focuses on different forensics linguistics aspects: author identification Kestemont *et al.* (2018), profiling Rangel *et al.* (2018), and obfuscation Hagen *et al.* (2018), whose aims, given a document, are respectively: to infer who wrote it, what are its author's demographic traits and to hide it.

When organising evaluation tasks, textual data (as well as multimedia one) should be labelled with information related to its content (e.g., irony, sentiment) or its author (e.g., gender, age, personality traits). In some cases, these data may be considered personal data (or personal data can be inferred from them). Therefore, the General Data Protection Regulation (GDPR),[3] the European regulation concerning the protection of individuals from the inappropriate use of their personal data Voigt and Von dem Bussche (2017), is of direct application. This regulation contains 99 (very restrictive Zarsky (2016)) articles, albeit we will focus only on those which directly apply to the scientific activities of organising evaluation tasks.

Likewise, before the download and reuse of data in the aforementioned evaluation tasks, the particular terms of use of the social platform from where the data is going to be collected must be taken into account. We will use Twitter as case study to illustrate its conditions, being the microblog platform that in most cases we used to collect the data for the PAN author profiling tasks, even though the presented methodology can (and must) be applied also to other platforms such as Facebook. In particular, the following should be considered when dealing with data for evaluation purposes:

- General Data Protection Regulation, mandatory when working with personal data (or from which personal data can be inferred) in/from/of the European Union.
- Particular terms of use of the specific social platform from where the data is collected. Concretely:
  - Legal base that allows the data treatment.
  - Permitted and prohibited behaviours related to collection, use and distribution of data.
  - The way to share and distribute data.
  - Other considerations that might reinforce the legal base for its utilisation.

The rest of the paper is structured as follows. In Section 2, we describe the legal framework of GDPR, focusing on the articles that directly apply to the organisation of evaluation tasks[4]. In Section 3 we illustrate the methodology to follow when organising

an evaluation task. In Section 4 we present a case study. Concretely, we explain how we have applied the proposed methodology for the organisation of the Author Profiling task at PAN, showing the particularities of the social platform Twitter. In Section 5, we overview the created author profiling corpora and how GDPR was applied. Finally, in Section 6 we draw the conclusions of this study.

Overview of the General Data Protection Regulation

The General Data Protection Regulation was approved on April 27, 2016 with the aim at protecting natural persons with regard to the processing of personal data and on the free movement of such data. The GDPR is applicable as of May 25th, 2018 in all member states to harmonize data privacy laws across Europe.[5]

It is noteworthy that the GDPR has been developed on the basis of the principle of proactive responsibility. This principle assumes *the necessity that the responsible of the treatment applies technical and organisational measures to guarantee and demonstrate that the data treatment is according to the Regulation.*

This principle requires a conscious, diligent and proactive attitude regarding the processing of personal data. It requires to analyse what data is treated, for what purpose and what type of treatment operations are carried out. To *guarantee* and *demonstrate* mean that it must be explicitly determined how the required measures will be implemented, that these measures are adequate to comply with the Regulation and that this fact can be demonstrated to all the interested parties and to the supervisory authorities.

Bearing in mind with this principle, from the 99 articles that make up the legal text, we focus only on those that directly affect the organisation of evaluation tasks.

Article 4. Definitions

This article defines the needed concepts for the purpose of the Regulation. The first point defines personal data as any information that identifies or can be used to identify a natural person. This definition is of high interest since it determines whether the Regulation must be complied.

> 1. 'personal data' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.

Article 6. Lawfulness of processing (legal base)

One of the keys of the law is to identify the legal base that allows the personal data treatment. In the case of evaluation tasks, the only possibility is defined in Article 6 (1) a).

> 1.a) the data subject has given consent to the processing of his or her personal data for one or more specific purposes.

Article 7. Conditions for consent

If the legal base is the express consent of the subject, we should demonstrate such consent according to Article 7 (1).

> 1. Where processing is based on consent, the controller shall be able to demonstrate that the data subject has consented to processing of his or her personal data.

Article 8. Conditions applicable to child's consent in relation to information society services

This article regulates the conditions of consent when dealing with minors. For example, when a minor sign up in a social network, this article is mandatory.

> 1. Where point (a) of Article 6 (1) applies, in relation to the offer of information society services directly to a child, the processing of the personal data of a child shall be lawful where the child is at least 16 years old. Where the child is below the age of 16 years, such processing shall be lawful only if and to the extent that consent is given or authorised by the holder of parental responsibility over the child.
>
> Member States may provide by law for a lower age for those purposes provided that such lower age is not below 13 years.

Article 9. Treatment of special categories of personal data

Article 9 (1) refers to personal data "*revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation*" and says that "*shall be prohibited.*"

However, in (2) there are some exceptions that may apply:

> e) the treatment refers to personal data that the interested party has made manifestly public.
>
> j) the treatment is necessary for the purposes of archiving in the public interest, scientific or historical research purposes, or statistical purposes, in accordance with Article 89, paragraph 1 [...]

Article 17. Right of suppression

This article refers to the right of users to delete their data at anytime. Nevertheless, there is an exception to this rule that may apply:

> 3. d) It will not apply when the treatment is necessary for the purposes of archiving in the public interest, scientific or historical research purposes, or statistical purposes, in accordance with Article 89, paragraph 1 [...]

Article 22. Automated individual decision-making, including profiling

This article is the most controversial one since it prohibits the automated profiling of users (one of the aims of forensic linguistics). Nonetheless, there is a nuance that may allow the organisation of evaluation tasks since they do not produce legal effects:

> 1. The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.

Article 24. Responsibility of the controller

This article (and subsequent Arts. 25, 30, 32, and 89) regulates the principle of proactive responsibility since we not only must apply technical and organisational measures, but also to be able to demonstrate them:

> 1. Taking into account the nature, scope, context and purposes of processing as well as the risks of varying likelihood and severity for the rights and freedoms of natural persons, the controller shall implement appropriate technical and organisational measures to ensure and to be able to demonstrate that processing is performed in accordance with this Regulation. Those measures shall be reviewed and updated where necessary.

Article 25. Data protection by design and by default

Two principles should be followed (*data minimisation* and *pseudonymisation*) to difficult, among others, the inverse identification of people:

> 1. Taking into account the state of the art, the cost of implementation and the nature, scope, context and purposes of processing as well as the risks of varying likelihood and severity for rights and freedoms of natural persons posed by the processing, the controller shall, both at the time of the determination of the means for processing and at the time of the processing itself, implement appropriate technical and organisational measures, such as pseudonymisation, which are designed to implement data-protection principles, such as data minimisation, in an effective manner and to integrate the necessary safeguards into the processing in order to meet the requirements of this Regulation and protect the rights of data subjects.
> 2. The controller shall implement appropriate technical and organisational measures for ensuring that, by default, only personal data which are necessary for each specific purpose of the processing are processed. That obligation applies to the amount of personal data collected, the extent of their processing, the period of their storage and their accessibility. In particular, such measures shall ensure that by default personal data are not made accessible without the individual's intervention to an indefinite number of natural persons.

Article 30. Records of processing activities

An organisational measure to be taken into account is to record all the processing activities, such as for example when data is released to the research community. In this article is also described the information that should be registered:

> 1. Each controller and, where applicable, the controller's representative, shall maintain a record of processing activities under its responsibility. That record shall contain all of the following information:
> a) the name and contact details of the controller and, where applicable, the joint controller, the controller's representative and the data protection officer;
> b) the purposes of the processing;
> c) a description of the categories of data subjects and of the categories of personal data;
> d) the categories of recipients to whom the personal data have been or will be disclosed including recipients in third countries or international organisations;
> e) where applicable, transfers of personal data to a third country or an international organisation, including the identification of that third country or international organisation and, in the case of transfers referred to in the second subparagraph of Article 49(1), the documentation of suitable safeguards;
> f) where possible, the envisaged time limits for erasure of the different categories of data;
> e) where possible, a general description of the technical and organisational security measures referred to in Article 32(1).

Article 32. Security of processing

Besides *data minimisation* and *pseudonymisation* described in Article 25, data processing must be ensured with technical measures such as *encryption*:

> 1. Taking into account the state of the art, the costs of implementation and the nature, scope, context and purposes of processing as well as the risk of varying likelihood and severity for the rights and freedoms of natural persons, the controller and the processor shall implement appropriate technical and organisational measures to ensure a level of security appropriate to the risk, including inter alia as appropriate:
> a) the pseudonymisation and encryption of personal data;

Article 89. Safeguards and derogations relating to processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes

Although Article 89 describes safeguards to be implemented, it is worth to mention some derogations that may apply in case of scientific research purposes, such as the organisation of evaluation tasks:

> 1. Processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes, shall be subject to appropriate safeguards, in accordance with this Regulation, for the rights and freedoms of the data subject. Those safeguards shall ensure that technical and organisational measures are in place in particular in order to ensure respect for the principle of data minimisation. Those measures may include pseudonymisation provided that those purposes can be fulfilled in that manner. Where those purposes can be fulfilled by further processing which does not permit or no longer permits the identification of data subjects, those purposes shall be fulfilled in that manner.
> 2. Where personal data are processed for scientific or historical research purposes or statistical purposes, Union or Member State law may provide for derogations from the rights referred to in Articles 15, 16, 18 and 21 subject to the conditions and safeguards referred to in paragraph 1 of this Article in so far as such rights are likely to render impossible or seriously impair the achievement of the specific purposes, and such derogations are necessary for the fulfillment of those purposes.

Methodology

In this section we propose a methodology to follow when organising evaluation tasks in order to ensure that GDPR, as well as the platform particular rules, are fulfilled when collecting, processing and distributing corpora *that contain personal data, or may contain identifiable personal data,* for scientific research purposes. It is noticeable the need to determine whether the corpora contain personal data as defined in GDPR Article 4 in order to apply (or not) the Regulation. The proposed methodology follows the schema represented in Figure 1 and it can be summarised in the following steps:

- To identify the legal base and to be able to demonstrate it.
- To consider special cases such as minors, special categories of data, or automatic profiling, whether some of them apply.
- To implement appropriate technical and organisational measures to ensure data protection.
- To distribute data according with both the social platform rules and the right of suppression.

- To record all the activities carried out with the data.
- Other considerations that may reinforce the legal framework to use the data in evaluation tasks.

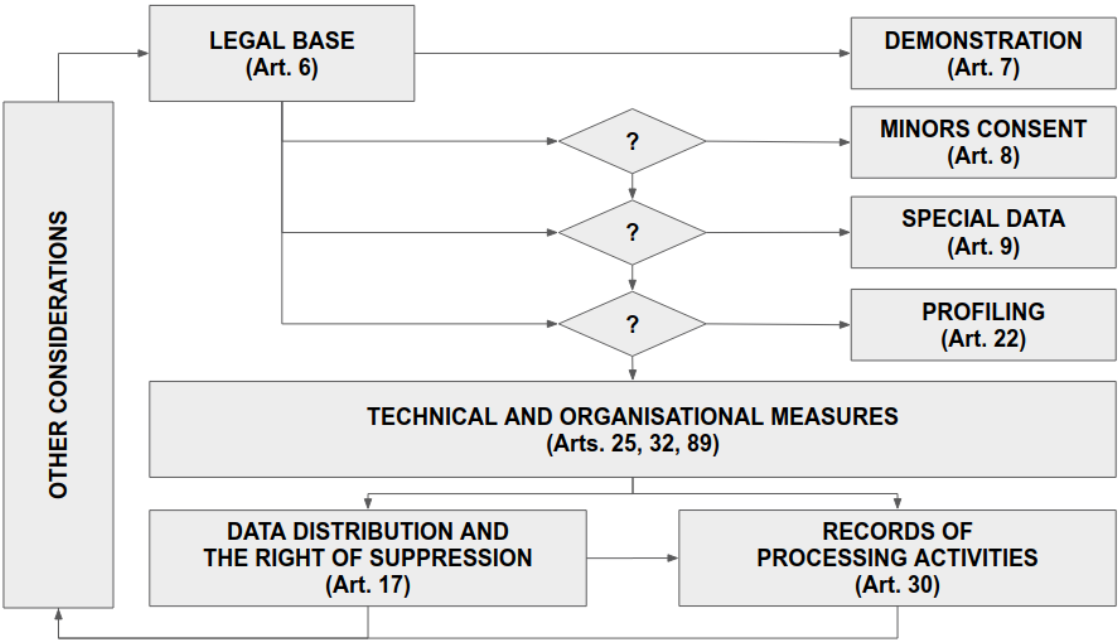

Figure 1. Methodology to accomplish GDPR when organising evaluation tasks, including automatic profiling.

The legal base and its demonstration

Following the GDPR principle of proactive responsibility, the first step is to determine the legal base that allows the use of the data in the evaluation task (Art. 6), as well as its demonstration (Art. 7). In case of evaluation tasks where the data is collected from social platforms the unique legal base that applies is *subject consent*. In such a case, it shall be demonstrated that the subjects gave their consent to use their data, in particular to use their data in evaluation tasks. This consent should be found in the terms of service of the social platform where the data is collected from. If this consent cannot be found, the data should not be used in the evaluation task.

Special cases: minors, special categories, automatic profiling

More attention should be paid when dealing with special cases such as minors (Art. 8), special categories of data (Art. 9), or automatic profiling (Art. 10). With respect to minors, the European Commission fixes the minimum age to consent at 16, albeit it allows the Member States to reduce that age as much as 13. In such cases, the consent shall be given by the legal guardian of the minor. Whether data from minors may be collected and used in the evaluation task, the organisers must ensure that the consent by the legal guardian was given. To do so the organisers should investigate how the social platform deals with minors and how it obtains the appropriate legal consent.

According to GDPR Article 9, the processing of special categories of personal data *shall be prohibited*. The first step is to determine whether the evaluation task needs or uses this kind of data. If it is needed, there are two exceptions (Section 2 of the Article) to the rule that may allow the use of this kind of data:

- *j)* Data is used for *specific research purposes*, which is the main purpose of evaluation tasks.
- *e)* Data *made manifestly public*. For each kind of special data, the organisers must ensure that the user made it manifestly public (e.g., giving public permissions to the reported birthday).

Automatic profiling is prohibited according to GDPR Article 22, but there is a nuance that may allow it in case of *non-commercial research purposes*.

Technical and organisational measures

GDPR urges to implement adequate technical and organisational measures to ensure that the data is secured, and to be able to demonstrate them. It should be followed the principles of *data minimisation* (Art. 25.1) and the *difficult to inverse identification of people* (Art. 89.1). To accomplish these principles, measures such as *encryption* (Art 32.1) and *pseudonymisation* (Art. 25.1) should be implemented.

Data distribution and the right of suppression

Data distribution must follow both the social platform rules and GDPR. In this regard and by applying the aforementioned technical and organisational measures, data should be released to the community encrypted and avoiding extra information that may allow the identification of personal data. This must be combined with the particular terms of service of the social platform which sometimes requires the release only of unique identifiers (e.g., Twitter). This situation should be analysed in each particular case.

In a similar vein, the right of suppression (Art. 17) allows users to delete their data at anytime. Deletion of the original data in the social platform should imply the automatic deletion of the data in the dataset of the evaluation task, albeit it might difficult the research activity (Art. 89.2) and the reproducibility of the experiments (Art. 17.3.d).

Records of processing activities

According to GDPR Article 30 all processing activities must be recorded. A special case is when data is released to third parties (e.g., to the participants of the evaluation task). It is imperative to implement the following measures:

- To register, at least, who is given access to the data, when, by whom, and what data in particular. It is recommendable to maintain a shared record (e.g., Google Sheet) with all the organisers, although only one of them should be the responsible to modify the register.
- To inform the researchers who receive the data that the only allowed purpose is *non-commercial scientific research*.

Other considerations

Depending on the task and the data to be used, other considerations may be extracted from the GDPR or the social platform terms of service. For example, if working with special categories of personal data such as (presumed) pedophiles that should not be available publicly, it may activate the *public interest* section in many GDPR articles that reinforce the legal base to use this data in the evaluation task.

Case Study: Author Profiling shared task at PAN

Since 2013 we have been organising at PAN an evaluation task on Author Profiling Rangel *et al.* (2013, b,a,c, 2017, 2018). With the exception of some years where data was collected also from other sources, we have mainly focused on Twitter data due to its availability, freedom of their users to express themselves and its idiosyncrasy for forensic linguistics.

In this section we describe how the proposed methodology has been applied to the organisation of the aforementioned evaluation campaigns, emphasising specific particularities of the task (e.g., dealing with special categories of data such as users personality traits or (presumed) pedophiles) and the social media platform (Twitter). Regarding the latter, besides GDPR we must fulfil the particular terms of the social media platform the data is collected from. In case of Twitter this information can be found in:

- Twitter Terms of Service[6], where the legal base for the data treatment is provided.
- Twitter Developer Policy[7], that indicates how data can be shared and distributed.
- Twitter Rules[8], that manifests prohibited behaviours for Twitter users, such as harassment or incitement to hatred, that allow us to make other considerations that reinforce our legal arguments.

To obtain the legal base and to be able to demonstrate it

As previously mentioned, according to GDPR Article 6, the unique legal base that applies is the *subject consent*. Furthermore, according to GDPR Article 7 we must be able to demonstrate that the subject consented. From the Twitter Terms of Service we can extract the needed legal base and its demonstration since Twitter is ensuring that the users consent, among others, the use of their data by third parties. Concretely, in Article *3. Content of the services*, in *Your rights and grants of rights in the contents*, Twitter users agree with the following (this must be accepted when a Twitter account is created):

> By submitting, posting or displaying Content on or through the Services, you grant us a worldwide, non-exclusive, royalty-free license (with the right to sublicense) to use, copy, reproduce, process, adapt, modify, publish, transmit, display and distribute such Content in any and all media or distribution methods (now known or later developed). This license authorizes us to make your Content available to the rest of the world and to let others do the same. You agree that this license includes the right for Twitter to provide, promote, and improve the Services and to make Content submitted to or through the Services available to other companies, organizations or individuals for the syndication, broadcast, distribution, promotion or publication of such Content on other media and services, subject to our terms and conditions for such Content use. Such additional uses by Twitter, or other companies, organizations or individuals, may be made with no compensation paid to you with respect to the Content that you submit, post, transmit or otherwise make available through the Services.

The consent in case of minors

When organising evaluation tasks with Twitter data we should take into account the possibility of using personal data from minors. In this regard and according to GDPR Article 8 regarding the consent of minors, explicitly this responsibility is derived to the holder of the parental responsibility.

In the Twitter Terms of Service, in Article *1. Who may use the services?*, it is stipulated that:

> [...] you must be at least 13 years [...]

GDPR stipulates the minimum age at 16, even though it allows the Member States to lower it:

> Member States may provide by law for a lower age for those purposes provided that such lower age is not below 13 years.

Hence, depending on the Member State, this age can be ranged between 13, that Twitter requires, and 16, required by the Regulation. In such cases, Twitter should be obligated to obtain the consent from the legal guardian of the minor, according to the aforementioned article. For example, in the adaptation of the GDPR that is being processed in Spain, in the Report of the Presentation on the Organic Law Project on Personal Data Protection 121/000013[9], of October 9, 2018, in its Article 7 on the Consent of minors, in its section 1, stipulates:

> 1. The treatment of personal data of a minor may only be based on his consent when he is older than 14 years.

In this case, when the national law is effective, if the minor is between 13 and 14, Twitter shall ensure that the consent to use its services was given by the holder of the parental responsibility at the moment of the account creation. In conclusion, this nuance reinforces the argument of the legal base (the consent), no matter the data might come from minors.

Dealing with special categories of personal data

According to GDPR Article 9 (1), the processing of special categories of personal data *shall be prohibited.* In linguistic forensics evaluation tasks we use to work with some of these special categories. For instance, when working on author profiling (e.g., personality traits) or stance detection (e.g., stance in favour or against some political matter). However, both exceptions e) (*data made manifestly public*) and j) (*scientific research purposes*) from Section 2 of the above article allow us to work with these kinds of data. Furthermore, Twitter Terms of Service, Section *3. Content of the services* reinforces the aforementioned exception e):

> You are responsible for your use of the Services and for any Content you provide, including compliance with applicable laws, rules, and regulations. You should only provide Content that you are comfortable sharing with others.

Automatic profiling

According to GDPR Article 22 profiling is prohibited. However, as we showed previously, there is a nuance that may allow our scientific activities since they do not produce legal or similarly significantly effects. Due to that, we inform researchers that the only allowed processing is for non-commercial research purposes (see Figure 2).

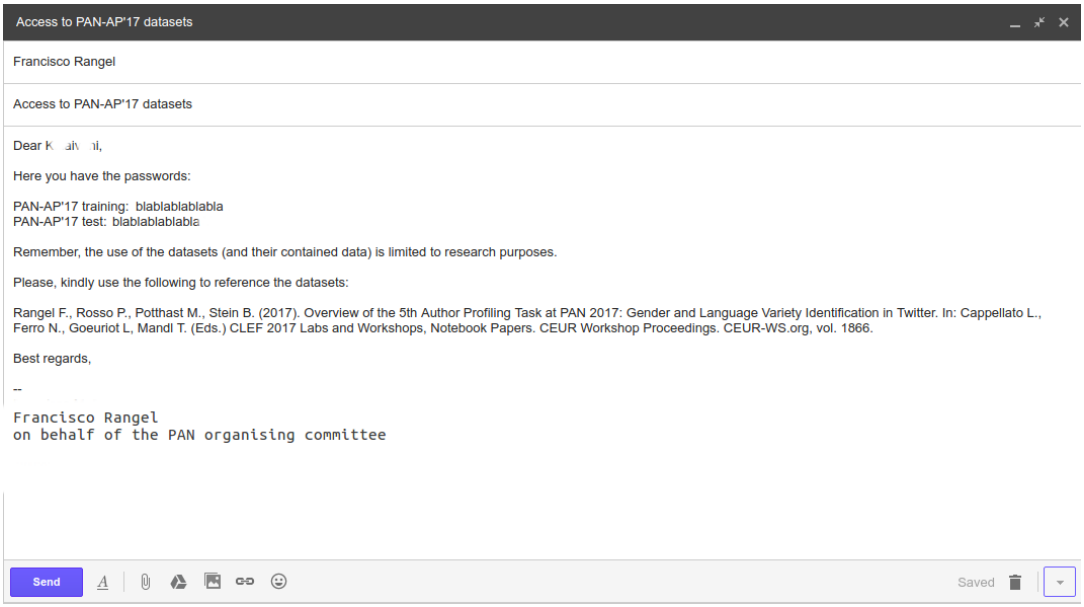

Figure 2. Email to give access to the dataset.

Technical and organisational measures

According to GDPR Articles 24, 25, 32 and 89, it is mandatory to implement the appropriate technical and organisational measures to ensure and be able to demonstrate that the data is secured. Concretely, we have implemented the following measures:

- To ensure that data is *pseudonymised* (Arts. 25, 32, and 89), we remove user mentions and other personal information (e.g., replacing mentions by @mention)[10].
- To ensure *data minimisation* principle (Arts. 25 and 89), we only distribute texts written by the authors and the corresponding labels (e.g., gender, age, etc.). An example of data format is shown in Figure 3.

```
<author id="1a9b3eacde983317d2e6b906232fbf06" lang="en" variety="new zealand" gender="female">
    <documents>
        <document><![CDATA[It looks like it is going to be ok after all ..or is it? https://t.co/8BpW6qun2r]]></document>
        <document><![CDATA[Setting up a giant marquee in the sun. Maybe I should switch… https://t.co/ku4dKg62Ua]]></document>
        <document><![CDATA[Just when I am about to go to sleep :/ #eqnz not cool at all]]></document>
        <document><![CDATA[#PJHarvey was bloody amazing tonight! https://t.co/5PM4zLZfIr]]></document>
        <document><![CDATA[@sue_fg what a way to close a brilliant show. Still buzzing..]]></document>

        ...

    </documents>
</author>
```

Figure 3. Data minimisation principle distributing only textual contents and labels.

- To ensure that data cannot be accessed freely without intervention (Art. 25 (2) and 32), data:
  - is *encrypted* when stored and distributed. We compress it with a 16 random generated characters.
  - is distributed only to known people that contacted us to ask for the password (as shown in the next subsection, this allows us to track processing activities).

Data distribution and the right of suppression

In the Twitter Developer Policy, in *F. Be a Good Partner to Twitter* is explicitly said how we should distribute the tweets. According to the original text shown below, Twitter only allows the distribution of its contents (tweets, users or direct messages) via its unique identifier (ID):

> 2. If you provide Twitter Content to third parties, including downloadable datasets of Twitter Content or an API that returns Twitter Content, you will only distribute or allow download of Tweet IDs, Direct Message IDs, and/or User IDs.

However, there are some exceptions that may favour and ease the organisation of evaluation tasks. Basically, it can be downloaded other information than IDs via non-automated means, as well as it can be surpassed both the distribution limit and the storage time limit for non-commercial research purposes:

> a) You may, however, provide export via non-automated means (e.g., download of spreadsheets or PDF files, or use of a "save as" button) of up to 50,000 public Tweet Objects and/or User Objects per user of your Service, per day.
> b.i) You may not distribute more than 1,500,000 Tweet IDs to any entity (inclusive of multiple individual users associated with a single entity) within any given 30 day period, unless you are doing so on behalf of an academic institution and for the sole purpose of non-commercial research or you have received the express written permission of Twitter.
> b.ii) You may not distribute Tweet IDs for the purposes of (a) enabling any entity to store and analyze Tweets for a period exceeding 30 days unless you are doing so on behalf of an academic institution and for the sole purpose of non-commercial research or you have received the express written permission of Twitter, or (b) enabling any entity to circumvent any other limitations or restrictions on the distribution of Twitter Content as contained in this Policy, the Twitter Developer Agreement, or any other agreement with Twitter.

GDPR Article 17 refers to the right of users to suppress their data. In this regard, Twitter users can delete their account or some of their tweets, and they also should be deleted from the datasets. This will occur if Twitter general rule of distributing only IDs is followed. However, GDPR Article 17 contains the exception (3) d) that allows to not applying the right of suppression in case of scientific research purposes. We can argue in favour of providing pseudonymised texts than tweets IDs taking into account the exception a) from the Article 2 of the Twitter Developer Policies, as well as GDPR Articles 17, 25, 32 and 89, in order to:

- maintain the reproducibility of the experiments, according to Article 17 (3) d).
- ease the research activity, according to Article 89 (2).
- difficult the inverse identification of people, according to Article 89 (1).
- follow the principle of data minimisation, according to Article 25 (1).
- apply technical and organisational measures such as encryption and pseudonymisation, according to Articles 32 (1) and 25 (1) respectively.

Records of processing activities

GDPR Article 30 compels to maintain a record of all processing activities regarding personal data, for example, when the data is distributed to a research team. At the PAN lab,

we maintain a list with all the people we send the data to, as well as we inform them about the only allowed purpose for the data (non-commercial research purposes).

	A	B	C	D	E
1	DATE	GIVEN BY	GIVEN TO	DATASET	ACTION
2	08/08/2018	Francisco Rangel	[illegible]	PAN-AP'18	Password for the test set
3	09/08/2018	Francisco Rangel	[illegible]	RUSPROFILING	Password for the whole dataset
4	18/08/2018	Martin Potthast	[illegible]	PAN-AP'18	Password for the training dataset
5	03/09/2018	Francisco Rangel	[illegible]	PR-SOCO	Password for the dataset
6	03/09/2018	Francisco Rangel	[illegible]	PAN-AP'17	Password for the whole dataset
7	03/09/2018	Francisco Rangel	[illegible]	PAN-AP'18	Password for the test dataset
8	03/09/2018	Francisco Rangel	[illegible]	PAN-AP'18	Password for the whole dataset
9	03/09/2018	Francisco Rangel	[illegible]	PAN-AP'15	Password for the whole dataset
10	03/09/2018	Francisco Rangel	[illegible]	PAN-AP'15	Password for the whole dataset
11	03/09/2018	Francisco Rangel	[illegible]	PAN-AP'17	Password for the whole dataset
12	03/09/2018	Francisco Rangel	[illegible]	PAN-AP'18	Password for the whole dataset
13	04/09/2018	Francisco Rangel	[illegible]	PAN-AP'18	Password for the test dataset
14	12/09/2018	Francisco Rangel	[illegible]	PAN-AP'17	Password for the test set

Figure 4. Excel sheet recording all processing activities regarding PAN datasets.

In Figure 4 an example of this record is shown in the form of an Excel sheet. Similarly, in Figure 2 we show an example of the informative email sent to the requester of the data. In this email we provide with the dataset passwords, inform about the unique allowed purpose of its use and kindly request the researcher to cite the overview paper where the dataset is described.

Other considerations

In the Authorship Attribution task at PAN 2012[11] Inches and Crestani (2012), a subtask on Sexual Predator Identification was organised. In the Author Profiling task at PAN 2013[12] Rangel *et al.* (2013) a subset of the previous data was also included. At present, we are organising the SemEval 2019 Shared Task 5 on Multilingual detection of hate speech against immigrants and women in Twitter (hatEval)[13]. In all these cases we work with very special categories of data (namely (presumed) pedophiles, misogynists, and racists). Twitter Rules do not allow users to behave abusively, such as for example sharing abusive, hateful or unwanted sexual contents. Twitter defines abusive behaviour as:

> Abuse: You may not engage in the targeted harassment of someone, or incite other people to do so. We consider abusive behavior an attempt to harass, intimidate, or silence someone else's voice.
> Unwanted sexual advances: You may not direct abuse at someone by sending unwanted sexual content, objectifying them in a sexually explicit manner, or otherwise engaging in sexual misconduct.
> Hateful conduct: You may not promote violence against, threaten, or harass other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. Read more about our hateful conduct policy.

In case there are tweets containing this kind of abusive contents because they have not been deleted by Twitter, according to GDPR Article 6 (1) e), if we try to identify them we would be working for the *public interest*:

> e) processing is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller;

Author Profiling Corpora

As mentioned before, we have been organising the Author Profiling task at PAN forensic linguistics Lab from 2013, both at CLEF[14] (Conferences and Labs of the Evaluation Forum) and FIRE[15] (Forum for Information Retrieval Evaluation). Every year we focus on different aspects of the authors (e.g., gender, age, personality traits, language variety) as well as on different languages (e.g., Arabic, Dutch, English, Italian, Portuguese, Spanish, Russian, or even computer languages such as Java). In this section we describe each of these corpora and how the GDPR was applied when created, processed and distributed (a summary can be seen in Table 1).

CORPUS	PERSONAL DATA	CONSENT	MINORS	SPECIAL CAT.	MEASURES DM	MEASURES EN	MEASURES PS	DISTRIB.
PAN AT CLEF								
PAN-AP'13	✗	✗	✓	✓	✓	✗	✗	Text + Labels
PAN-AP'14	✓	?	✓	✗	✓	✗	✗	Text + Labels
PAN-AP'15	✓	✓	✗	✓	✗	✓	✗	ID + Labels
PAN-AP'16	✓	?	✗	✗	✓	✗	✗	Text + Labels
PAN-AP'17	✓	✓	?	✗	✓	✓	✗	Text + Labels
PAN-AP'18	✓	✓	?	✗	✓	✓	✓	Text + Image + Labels
PAN AT FIRE								
RusProf'17	✓	✓	?	✗	✓	✓	✗	Text + Labels
PR-SOCO'16	✗	✓	✗	✓	✓	✓	✓	Text + Labels

LEGEND			
Data Min.	YES	✓	ID
ENcryption	NO	✗	Text
PSeudonym.	UNKNOWN	?	Image
			Labels

Table 1. Summary with the GDPR measures applied to the different Author Profiling corpora, identified in the first column. The second column reports whether the corpus may contain personal data and, in such a case, if the users consented. In the third and fourth columns the occurrence of minors and special categories of data are represented respectively. Columns five to seven show the technical and organisational measures applied, whereas column eight indicates the type of data distributed within the corpus. A legend is given at the bottom of the table.

Age and Gender Identification in Social Media (PAN-AP'13 at CLEF)

The focus of the 2013 evaluation task was on age and gender identification in social media. We tried to emulate a realistic big data scenario looking for open and public online repositories such as Netlog[16] with posts labelled with author demographics (gender and age). Following pioneer investigations Schler *et al.* (2006), we considered three age groups: 10s (13-17), 20s (23-27), and 30s (33-47). We also incorporated a small number of samples of adult-adult conversations about sex together with conversations of sexual predators Inches and Crestani (2012) with the aim of investigating the robustness of

the state-of-the-art of age identification systems to unveil the age of sexual predators (usually pretending to be minors). In Table 3 we show the statistics of the English and Spanish corpora[17]. The corpus was balanced by gender and imbalanced by age group. More information can be found in the evaluation task overview paper Rangel *et al.* (2013).

		ENGLISH		SPANISH	
Age	**Gender**	**No. of Authors**		**No. of Authors**	
		Training	Test	Training	Test
10s	male	8 600	888	1 250	144
	female	8 600	888	1 250	144
20s	male	(72) 42 828	(32) 4 576	21 300	2 304
	female	(25) 42 875	(10) 4 598	21 300	2 304
30s	male	(92) 66 708	(40) 7 184	15 400	1 632
	female	66 800	7 224	15 400	1 632
Σ		236 600	25 440	75 900	8 160

Table 2. Distribution of the number of authors per class in PAN-AP'13 corpus.

Data were collected from the Netlog social platform that is no longer available. Hence, personal information cannot be inferred from the contents distributed in the corpus and, therefore, the GDPR does not apply. The corpus contains texts written by minors in the range of 10s (13-17), and texts from users labelled as sexual predators that can be considered special categories of data. We applied data minimisation by distributing only texts and labels corresponding to the author's age and gender. We did not encrypted data since the information was publicly available. Moreover, we did not applied pseudonymisation because we considered mentions to other people as significant for the task (however the sexual predators subset is anonymised). The first row of Table 1 summarises the described measures.

Multi-Genre Age and Gender Identification (PAN-AP'14 at CLEF)

The aim of the 2014 evaluation task was investigating how the author profiling approaches would perform on different genres: social media, blogs, Twitter and hotel reviews. The corpus covers English and Spanish languages (see Table 5), except in case of hotel reviews that are in English. That year, age ranges considered the following groups: 18-24, 25-34, 35-49, 50-64, and 65+. More information about the collection of the corpus can be found in the overview paper of the evaluation task Rangel *et al.* (b).

		ENGLISH		SPANISH	
Age	**Gender**	**No. of Authors**		**No. of Authors**	
		Training	Test	Training	Test
10s	male	8 600	888	1 250	144
	female	8 600	888	1 250	144
20s	male	(72) 42 828	(32) 4 576	21 300	2 304
	female	(25) 42 875	(10) 4 598	21 300	2 304
30s	male	(92) 66 708	(40) 7 184	15 400	1 632
	female	66 800	7 224	15 400	1 632
Σ		236 600	25 440	75 900	8 160

Table 3. Distribution of the number of authors per class in PAN-AP'13 corpus.

As there are several social media, we must determine whether each of them may contain personal data. The case of social media was discussed previously, and in case of blogs, personal data should not be inferred from contents unless the users explicitly published them. Thus, the GDPR does not apply for these social media.

In case of Twitter or reviews, personal data can be inferred from the contents and therefore they may contain personal data as defined in the Article 4 of GDPR. Due to the fact that in 2014 GDPR did not exist, the explicit consent was not mandatory and we cannot know if these platforms required it at that time. Nowadays, the social platforms must obtain the consent of the users in case they did not already give it. The users can revoke this consent or exercise the right of suppression described in Article 17. In such cases, we shall appeal to the exception 3.d) of the same article to maintain the data for scientific research purposes. In any case, we do not know whether the consent was given.

The corpus contains texts written by minors in the range of 10s (13-17) and it does not contain special categories of data. We applied data minimisation by distributing only texts and labels with age and gender information. We did not encrypted data since it was publicly available, as well as we did not applied pseudonymisation because we considered mentions to other people as significant for the task. The described measures are summarised in the second row of Table 1.

Age, Gender and Personality Recognition in Twitter (PAN-AP'15 at CLEF)

The author profiling evaluation task at PAN 2015 focused on age, gender and personality recognition of Twitter users. The most widely theory in psychology to define personality is Five Factor Theory Costa and McCrae (1985, 2008). This theory defines five traits (OCEAN): openness to experience (O), conscientiousness (C), extroversion (E), agreeableness (A), and emotional stability / neuroticism (N). To annotate the data we created an online questionnaire asking for age, gender and personality traits following the BFI-10-test Rammstedt and John (2007). Personality scores were normalised between -0.5 and +0.5, and we used the following age groups: 18-24, 25-34, 35-49, 50+. Except for age, the corpus covers English, Spanish, Italian and Dutch. The corpus statistics are shown in Table 4 and more information can be found in the overview paper of the evaluation task Rangel *et al.* (a).

	Training				Test			
	EN	ES	IT	DU	EN	ES	IT	DU
Users	152	110	38	34	142	88	36	32
18-24	58	22			56	18		
25-34	60	56			58	44		
35-49	22	22			20	18		
50+	12	10			4 8	8		
Male	76	55	19	17	71	44	18	16
Female	76	55	19	17	71	44	18	16
E (mean)	0.16	0.18	0.17	0.24	0.17	0.16	0.15	0.24
S (mean)	0.14	0.07	0.20	0.21	0.13	0.09	0.20	0.22
A (mean)	0.12	0.14	0.22	0.13	0.14	0.14	0.19	0.15
C (mean)	0.17	0.24	0.18	0.14	0.17	0.21	0.21	0.17
O (mean)	0.24	0.18	0.23	0.29	0.26	0.19	0.25	0.28

Table 4. Distribution of the number of authors per class in PAN-AP'15 corpus.

As Twitter users can be identified from their contents, the tweets should be considered as personal data. Although the Regulation should apply from 25 May 2018, in 2016 entered into force. Thus, we followed its Article 6 and requested the explicit consent of the users to process their data for research purposes. The users had to consent before filling out the aforementioned questionnaire.

This corpus does not contain data from minors since the lowest age is 18. It may be considered the existence of special categories of data regarding personality traits. We followed Twitter rule of distributing tweet IDs, thus we could not apply the data minimisation criteria nor the pseudonymisation. The applied measures are summarised in the third row of Table 1.

Cross-Genre Age and Gender Identification (PAN-AP'16 at CLEF)

In the 2016 evaluation task, we aimed at investigating the effect of the cross-genre evaluation: how the models perform when they are trained on one genre and evaluated on another different genre. In this regard, the training corpus was collected from Twitter for the three languages: Dutch, English, and Spanish. In case of Spanish and English, we merged the training and test sets from PAN-AP'14 Twitter corpus Rangel *et al.* (b), whilst in case of Dutch, the training corpus was mined as a precursor of TwiSty Verhoeven *et al.* (2016). The test corpus for English and Spanish was obtained from the test partition of the PAN-AP'14 blog subcorpus. Furthermore, as in previous years we provided with an early bird evaluation. However, unlike in previous years where early birds used a subset from the test set, this year we took advantage of this early evaluation to evaluate another genre. In concrete, early birds data in English and Spanish was collected from the social media subset of the PAN-AP'14 corpus. The test set (both early and final tests) for Dutch combined reviews from the CSI corpus Verhoeven and Daelemans (2014) and student essays. As shown in Table **??**, in case of Dutch only gender information is provided, whereas for English and Spanish the following age groups are covered: 18-24, 25-34, 35-

49, 50-64, 65+. More information about the corpora can be found in the overview paper of the evaluation task Rangel *et al.* (c).

PAN-AP'16 corpus was created from PAN-AP'14, thus what was discussed there it also applies here. The only exception is that there are no minors in 2016 corpus since the lowest age was increased to 18. A summary of measures can be seen in the fourth row of Table 1.

	ENGLISH								SPANISH					
	SocialMedia		Blog		Twitter		Reviews		SocialMedia		Blog		Twitter	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
18-24	1 550	680	6	10	20	12	360	148	330	150	4	4	12	4
25-34	2 098	900	60	24	88	56	1 000	400	426	180	26	12	42	26
35-49	2 246	980	54	32	130	58	1 000	400	324	138	42	26	86	46
50-64	1 838	790	23	10	60	26	1 000	400	160	70	12	10	32	12
65+	14	26	4	2	8	2	800	294	30	28	4	2	6	2
Σ	7 746	3 376	147	78	306	154	4 160	1 642	1 272	566	88	56	178	90

Table 5. Distribution of the number of authors per class in PAN-AP'14 corpus.

Gender and Language Variety Identification in Twitter (PAN-AP'17 at CLEF)

The focus of the 2017 evaluation task was on gender and language variety identification in Twitter. The corpus included four languages: Arabic, English, Portuguese and Spanish. We retrieved tweets geolocated in the capital cities where the target language variety is used. Unique users were selected and annotated with the corresponding variety. A dictionary with proper nouns was used to annotate the users' gender, as well as a manual inspection of their photo profiles was carried out to improve the annotation quality. Finally, for each user a hundred tweets were collected from her/his timeline. The corpus was divided into training/test in a 60/40 proportion, with 300 authors for training and 200 authors for test. The corresponding languages and varieties are shown in Table 6 along with the total number of authors for each subtask. More information about this corpus is available in the evaluation task overview paper Rangel *et al.* (2017).

(AR) Arabic	(EN) English	(ES) Spanish	(PT) Portuguese
Egypt	Australia	Argentina	Brazil
Gulf	Canada	Chile	Portugal
Levantine	Great Britain	Colombia	
Maghrebi	Ireland	Mexico	
	New Zealand	Peru	
	United States	Spain	
		Venezuela	
4,000	6,000	7,000	2,000

Table 6. Distribution of the number of authors per class in PAN-AP'17 corpus.

In the fifth row in Table 1 the applied GDPR measures when building and distributing the PAN-AP'17 corpus are summarised. As data was collected from Twitter, the consent was given to the social platform. It is not possible to know whether there are minors in the corpus because age was not verified. There are no data belonging to special categories since the unique provided label refers to users' gender. We applied data minimisation, since only texts and labels were distributed, as well as encryption since data was distributed compressed with password. We did not pseudonymised texts because nouns might contribute to the task.

Multi-Modal Gender Identification in Twitter (PAN-AP'18 at CLEF)

In 2018 we aimed to investigate the effect of multi-modal information on the gender identification task in Twitter. Multi-modal means that besides textual information, also images could be used. The corpus included three languages: Arabic, English and Spanish. This corpus was created as a subset of the PAN-AP'17 corpus. For each author, we collected all the images shared in her/his timeline. We discarded users who deleted their account as well as users with less than 10 images in their timeline. Each author contains exactly 100 tweets and 10 images. The corpus is completely balanced per gender and split in training/test sets as shown in Table 7.

	(AR) Arabic	(EN) English	(ES) Spanish	Total
Training	1,500	3,000	3,000	7,500
Test	1,000	1,900	2,200	5,100
Total	2,500	4,900	5,200	12,600

Table 7. Distribution of the number of authors per class in PAN-AP'18 corpus.

The sixth row of Table 1 summarises the applied measures. The only differences with PAN-AP'17 lie in the following: the distributed corpus contains also images, and this year we applied pseudonymisation by removing user mentions.

Cross-Genre Gender Identification in Russian (RUSPROFILING'17 PAN at FIRE)

Slavic languages have been less investigated from an author profiling standpoint and have never been addressed at PAN before. This task aimed at investigating gender identification in Russian from a cross-genre perspective. That is, we provided tweets as a training corpus and Facebook posts, online reviews, texts describing images or letters to a friend, as well as tweets as test corpus. In Table 8 a summary of the number of authors per genre is shown. More information on the corpus construction can be found in the overview paper of the evaluation task Litvinova *et al.* (2017).

RusProfiling'17 corpus contains data from different sources, even though we can group them into two types: social media platforms and students' essays. In case of social media platforms, as seen previously, personal data may be inferred from contents, coercing the application of the Regulation. In case of students' essays, although personal information should not be identifiable from their contents, the ease to obtain their consent worth it.

Table 1 summarises the GDPR measures that we applied to build and distribute the corpus. In case of social media platforms the consent was given when the account was

Dataset	Genre	Number of authors
Training	Twitter	600
Test	Essays	370
	Facebook	228
	Twitter	400
	Reviews	776
	Gender-imitated	94

Table 8. Distribution of the number of authors per genre in RusProfiling'17 corpus.

created, as well as in case of students' essays, the students gave their consent when participated. We cannot know whether there are minors in the data collected from social platforms since we did not verified the age, but we can ensure that there are no minors in the subsets of essays and gender-imitated since the authors were university students. There are no special categories of data because we only provided gender as labels. We applied both data minimisation and encryption to distribute only texts and gender labels, and we compressed the corpus with password. Pseudonymisation was not applied because mentions might contribute to the task.

Personality Recognition in SOurce COde (PR-SOCO'16 PAN at FIRE)

Finally, in the PR-SOCO evaluation task we aimed at investigating whether personality traits could be inferred from the way Java programming language is used by computer science students. Students were asked to write source code responding to some functional requirements of different programming tasks. In addition each student answered a Big Five personality test. The dataset consists of 2,492 source code programs written by 70 students (49 for training, 21 for test). The scores for the personality traits range between 20 and 80. More information about the corpus can be found in the overview paper of the evaluation task Rangel *et al.* (2016).

Despite the fact that natural persons should no be identifiable from the PR-SOCO'16 corpus, we applied GDPR measures because they were identifiable when collecting the data. Data was collected from students who explicitly expressed their consent. There are no minors since the subjects were university students of Computer Science, but the corpus does cover the special category of data regarding personality traits. We applied data minimisation, encryption and pseudonymisation: data minimisation since only source code and personality scores were distributed, encryption because the corpus was distributed compressed with password, and pseudonymisation in case some students incorporated personal nouns for instance in the source code comments. The corpus is distributed as plain text containing source code in Java language together with the labels corresponding to the five personality traits. In the last row of Table 1 we summarise the applied measures when the corpus was created and distributed.

Conclusions

The organisation of evaluation tasks allows the creation of a common framework for research, fostering comparability and reproducibility. Moreover, social data allows for investigating forensic linguistics aspects in a big data scenario. However, due to the implications that the release of the data may have on the privacy of people, the European

law for its protection must be contemplated. These norms are defined in the General Data Protection Regulation (GDPR) of April 27, 2016, as well as in the legal base of use of the particular social platform from where data are collected.

In this paper, we have proposed a methodology to follow when creating corpora for the organisation of an evaluation task. Firstly, we have described the GDPR articles that apply. For each article, we have highlighted the principal aspects as well as the plausible exceptions that may help in the organisation of the task. GDPR principle of proactive responsibility assumes that the responsible of the treatment, in this case the organiser of the evaluation task, applies technical and organisational measures to guarantee and demonstrate that the data treatment is according to the Regulation. Therefore, the first step is to identify (Art. 6) and demonstrate (Art. 7) the legal base for the treatment (i.e., subject consent). A special attention must be paid when dealing with special cases (Art. 8) (i.e., minors), special categories of data (Art. 9) (i.e., political options, religious of philosophical beliefs, sexual orientation, etc.), or the treatment implies (automatic) profiling (Art. 22). In such cases, the organiser must investigate whether the possible exceptions may apply (i.e., research purposes, data made manifestly public, etc.). Furthermore, the organiser must apply technical and organisational measures (Arts. 25, 32, 89) (i.e., data minimisation, encryption, pseudonymisation, etc.) to difficult the inverse identification of people. Finally, the organiser must distribute data according with both the social platform rules and the right of suppression (Art. 17) and to record all the processing activities carried out with the data (Art. 30). At least, to register who is given access to the data as well as to inform that the only allowed purpose is non-commercial scientific research.

With the aim at guiding researchers in the application of the GDPR to the organisation of shared tasks, we have presented a case study about the organisation of the forensic linguistic task on author profiling at the PAN Lab at CLEF, that we have been organising since 2013, showing how both GDPR and Twitter Terms of Service have been complied. Finally, we have described the different corpora created at PAN and how the Regulation was observed in these cases.

Acknowledgements

This article was made possible by NPRP grant 9-175-1-033 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

Notes

[1]https://pan.webis.de/

[2]http://www.clef-initiative.eu

[3]https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679

[4]We use italic when text is extracted from the legal source, and underline when we want to highlight something.

[5]It is worth to mention that the GDPR must be adapted to the local legislation of each Member State. This implies to translate the Regulation, at least, to 24 official languages. Furthermore, it shall be adapted to the cultural, social and legal particularities of each of the States Sosoni and Biel (2018).

[6]https://twitter.com/en/tos

[7]https://developer.twitter.com/en/developer-terms/policy.html

[8]https://help.twitter.com/en/rules-and-policies/twitter-rules

[9]http://www.congreso.es/public_oficiales/L12/CONG/BOCG/A/BOCG-12-A-13-1.PDF

[10]We do it in those cases where we consider that this information is not valuable for the specific task.

[11]https://pan.webis.de/clef12/pan12-web/author-identification.html

[12]https://pan.webis.de/clef13/pan13-web/author-profiling.html

[13]https://competitions.codalab.org/competitions/19935

[14]http://clef2018.clef-initiative.eu/

[15]http://fire.irsi.res.in

[16]https://www.netlog.com

[17]In the training part of the English collection, numbers inside parentheses for male 20s and 30s correspond to the number of samples of sexual predator conversations while numbers inside parenthesis for female 20s correspond to the adult-adult sexual conversation samples. The final collection includes samples from sexual predator conversations for male 20s and 30s, and samples from adult-adult conversations for female 20s.

References

Costa, P. T. and McCrae, R. R. (1985). *The NEO personality inventory: Manual, form S and form R*. Psychological Assessment Resources.

Costa, P. T. and McCrae, R. R. (2008). The revised neo personality inventory (neo-pi-r). *The SAGE handbook of personality theory and assessment*, 2, 179–198.

Coulthard, M., Johnson, A. and Wright, D. (2016). *An introduction to forensic linguistics: Language in evidence*. Routledge.

Hagen, M., Potthast, M. and Stein, B. (2018). Overview of the Author Obfuscation Task at PAN 2018. In *Working Notes Papers of the CLEF 2018 Evaluation Labs*, CEUR Workshop Proceedings: CLEF and CEUR-WS.org.

Inches, G. and Crestani, F. (2012). Overview of the International Sexual Predator Identification Competition at PAN-2012. In P. Forner, J. Karlgren and C. Womser-Hacker, Eds., *CLEF 2012 Evaluation Labs and Workshop – Working Notes Papers*.

Kestemont, M., Tschuggnall, M., Stamatatos, E., Daelemans, W., Specht, G., Stein, B. and Potthast, M. (2018). Overview of the Author Identification Task at PAN-2018: Cross-domain Authorship Attribution and Style Change Detection. In *Working Notes Papers of the CLEF 2018 Evaluation Labs*, CEUR Workshop Proceedings: CLEF and CEUR-WS.org.

Litvinova, T., Rangel, F., Rosso, P., Seredin, P. and Litvinova, O. (2017). Overview of the rusprofiling pan at fire track on cross-genre gender identification in russian. In *FIRE (Working Notes)*, 1–7.

Rammstedt, B. and John, O. (2007). Measuring personality in one minute or less: A 10 item short version of the big five inventory in english and german. In *J. Research in Personality*, 203–212.

Rangel, F., Celli, F., Rosso, P., Potthast, M., Stein, B. and Daelemans, W. In L. Cappellato, N. Ferro, G. Jones and E. San Juan, Eds., *CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers, 2015*.

Rangel, F., González, F., Restrepo, F., Montes, M. and Rosso, P. (2016). Pan@ fire: Overview of the pr-soco track on personality recognition in source code. In *Forum for Information Retrieval Evaluation*, 1–19: Springer.

Rangel, F., Rosso, P., Chugur, I., Potthast, M., Trenkmann, M., Stein, B., Verhoeven, B. and Daelemans, W. In L. Cappellato, N. Ferro, M. Halvey and W. Kraaij, Eds., *CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers, 2014*.

Rangel, F., Rosso, P., Koppel, M., Stamatatos, E. and Inches, G. (2013). Overview of the Author Profiling Task at PAN 2013. In P. Forner, R. Navigli and D. Tufis, Eds., *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers*.

Rangel, F., Rosso, P., Potthast, M. and Stein, B. (2017). Overview of the 5th Author Profiling Task at PAN 2017: Gender and Language Variety Identification in Twitter. In L. Cappellato, N. Ferro, L. Goeuriot and T. Mandl, Eds., *Working Notes Papers of the CLEF 2017 Evaluation Labs*, CEUR Workshop Proceedings: CLEF and CEUR-WS.org.

Rangel, F., Rosso, P., Verhoeven, B., Daelemans, W., Potthast, M. and Stein, B. In K. Balog, L. Cappellato, N. Ferro and C. Macdonald, Eds., *CLEF 2016 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings. CEUR-WS.org*.

Rangel, F., Rosso, P., y Gómez, M. M., Potthast, M. and Stein, B. (2018). Overview of the 6th author profiling task at pan 2018: Multimodal gender identification in twitter. In *CLEF 2018 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings. CEUR-WS.org*.

Schler, J., Koppel, M., Argamon, S. and Pennebaker, J. W. (2006). Effects of age and gender on blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, 199–205: AAAI.

Sosoni, V. and Biel, Ł. (2018). Eu legal culture and translation. *International Journal of Language & Law (JLL)*, 7.

Verhoeven, B. and Daelemans, W. (2014). Clips stylometry investigation (csi) corpus: a dutch corpus for the detection of age, gender, personality, sentiment and deception in text. In *9th International Conference on Language Resources and Evaluation (LREC 2014)*.

Verhoeven, B., Daelemans, W. and Plank, B. (2016). Twisty: a multilingual twitter stylometry corpus for gender and personality profiling. In *10th International Conference on Language Resources and Evaluation (LREC 2016)*.

Voigt, P. and Von dem Bussche, A. (2017). *The EU General Data Protection Regulation (GDPR)*, volume 18. Springer.

Zarsky, T. Z. (2016). Incompatible: The gdpr in the age of big data. *Seton Hall L. Rev.*, 47, 995.

Computational Forensic Linguistics: An Overview of Computational Applications in Forensic Contexts

Rui Sousa-Silva

Universidade do Porto, Portugal

Abstract. *The number of computational approaches to forensic linguistics has increased significantly over the last decades, as a result not only of increasing computer processing power, but also of the growing interest of computer scientists in natural language processing and in forensic applications. At the same time, forensic linguists faced the need to use computer resources in both their research and their casework – especially when dealing with large volumes of data. This article presents a brief, non-systematic survey of computational linguistics research in forensic contexts. Given the very large body of research conducted over the years, as well as the speed at which new research is regularly published, a systematic survey is virtually impossible. Therefore, this survey focuses on some of the studies that are relevant in the field of computational forensic linguistics. The research cited is discussed in relation to the aims and objectives of the linguistic analysis in forensic contexts, paying particular attention to both their potential and their limitations for forensic applications. The article ends with a discussion of future implications.*

Keywords: *Computational forensic linguistics, computational linguistics, authorship analysis, plagiarism, cybercrime.*

Resumo. *O recurso a abordagens computacionais na área da linguística forense aumentou drasticamente ao longo das últimas décadas, decorrente, não só ao aumento das capacidades de processamento dos computadores, mas também do interesse crescente de especialistas do ramo das ciências de computadores no processamento de linguagem natural e nas suas aplicações forenses. Simultaneamente, os linguistas forenses depararam-se com a necessidade de utilizar recursos informáticos, tanto nos seu trabalho de investigação, como nos seus casos de consultoria forense, sobretudo tratando-se do processamento de grandes volumes de dados. Este artigo apresenta uma revisão breve, não sistemática, da investigação científica em linguística computacional aplicada a contextos forenses. Tendo em conta o elevado volume de investigação publicada, bem como o ritmo acelerado de publicação nesta área, a realização de uma revisão bibliográfica sistemática é praticamente impossível. Por conseguinte, esta revisão foca alguns dos estudos mais relevantes na área da linguística forense computacional. Os estudos mencionados são discutidos no âmbito das metas e dos objetivos da análise linguística*

em contextos forenses, prestando-se atenção especialmente ao seu potencial e às suas limitações no tratamento de casos forenses. O artigo termina com uma discussão de algumas das implicações futuras da computação em aplicações forenses.

Palavras-chave: *Linguística forense computacional, linguística computacional, análise de autoria, plágio, cibercrime.*

Introduction

Forensic Linguistics has attracted significant attention ever since Svartvik (1968) published 'The Evans Statements: A Case for Forensic Linguistics' (Svartvik, 1968), not the least because the analysis reported by the author showed the true potential of linguistic analysis in forensic contexts. Since then research into – and the use of – forensic linguistics methods and techniques have multiplied, and so has the range of possible applications. Indeed, the three subareas identified by Forensic Linguistics in a broad sense – the written language of the law, interaction in legal contexts and language as evidence (Coulthard and Johnson, 2007; Coulthard and Sousa-Silva, 2016) – have been furthered, and extended to a plethora of other applications all over the world; the written language of the law came to include applications other than studying the complexity of legal language; interaction in legal contexts has significantly evolved, and now focuses on any kind of interaction in legal contexts – including attempts to identify the use of deceptive language (Gales, 2015), or ensure appropriate interpreting (Kredens, 2016; Ng, 2016); and language as evidence has gained a reputation of robustness and reliability, with further research on disputed meanings (Butters, 2012), the application of methods of authorship analysis in response to new needs (e.g. cybercriminal investigations), and an attempt to develop new theories, e.g. authorship synthesis (Grant and MacLeod, 2018).

It is perhaps as a result of the need to respond to new problems arising from the development of new information and communication technologies that language as evidence continues to be the most visible 'face' of Forensic Linguistics. The technological advances of the last decades have opened up new possibilities for forensic linguistic analysis: new forms of online interaction have required new forms of computer-mediated discourse analysis (Herring, 2004), and synchronous and immediate forms of communication such as the ones provided by online platforms have allowed users to communicate with virtually anyone based anywhere in the world and at any time from any mobile device, while replacing face-to-face with online interaction. At the same time, such technologies offered new anonymisation possibilities, both real and perceived. If, on the one hand, using stealth technologies and un-monitored, unsupervised public computers and networks grants users some level of real anonymity, on the other hand that anonymity is very often only perceived, rather than real. As such, although users can be easily identified – especially by law and order enforcement agents – the fact that they perceive themselves to remain anonymous behind the computer keyboard or the mobile phone display (e.g. by using fake profiles) encourages them to practice illegal acts that most people refrain from doing when face-to-face, including hate crimes, threats, libel and defamation, fraud, infringement of intellectual property, stalking, harassment and bullying.

Therefore, not only have such developments raised new (and exciting) challenges for forensic linguists, they have also demonstrated that new tools and techniques are required to handle data collection, processing and (linguistic) analysis quickly and ef-

ficiently. That is especially the case with large volumes of data, in which the linguist needs to face the 'big data' challenge, which consists of managing huge volumes of text. In fact, large volumes of data make it virtually impossible for linguists to manually process and analyse the data quickly and accurately. Therefore, they usually resort to the use of computational tools. Such an analysis can be heavily computational, i.e. it can be conducted with no or very little human intervention, or computer-assisted, in which computational tools and techniques are used as an aid to the manual analysis, e.g. in searching words or phrases, or comparing some textual elements against a reference corpus or tagging a text, among others.

The use of computational linguistics in forensic contexts has become so indispensable that it has given rise to the field of computational forensic linguistics. However, the meaning of the concept of computational forensic linguistics, like the concept of computational linguistics, is far from agreed, and people from different areas of expertise tend to conceive of the area differently. This article thus begins with a discussion of the concept and proposes a working definition to encompass work conducted by computer scientists on natural language processing, that is most helpful to forensic linguists. Subsequently, it presents a survey of methods and techniques that have contributed to forensic applications, including authorship analysis, plagiarism detection and disputed meanings. The article concludes with a discussion of both the potential and the limitations of computational analysis to argue that, although a purely computational analysis can be extremely valuable in forensic contexts, ultimately such an analysis can only be acceptable as an evidential or even an investigative tool when interpreted by a linguist.

Defining computational forensic linguistics

Woolls (2010: 576) defines computational forensic linguistics concisely as "a branch of computational linguistics" (CL), a discipline which Mitkov (2003: ix) had previously defined as "an interdisciplinary field concerned with the processing of language by computers". CL, although bearing a different name, originated in the 1940s with the work of Weaver (1955), especially based on his suggestion of the possibilities of machine translation. Over time, CL contributed to an array of applications across different usage domains, most of which can be potentially useful to forensic linguists, including machine translation, terminology, lexicography, information retrieval, information extraction, grammar checking, question answering, text summarisation, term extraction, text data mining, natural language interfaces, spoken dialogue systems, multimodal/multimedia systems, computer-aided language learning, multilingual online language processing, speech recognition, text-to-speech synthesis, corpora, phonological and morphological analysis, part of speech tagging, shallow parsing, word disambiguation, phrasal chunking, named entity recognition, text generation, user ratings and comments / reviews, and detection of fake news and hyperpartisanism.

However, CL did not develop uncontroversially over the years: as the field contemplates natural language (an object of study that is dear to linguistics) and its processing by computers (the role of computer science), CL has been amid a tension between linguists and computer scientists. From an early stage, computer scientists managed to show that computational approaches to linguistics had the potential to achieve more successful results than linguistic methods alone. They did so primarily by abandoning, at least in part, the overly fine-grained sets of rules that linguists have been arguing for, based especially on the work of Chomsky (1972); while linguists were focused on

language structure and use, computer scientists argued that more formalisms and more language models – and of a different nature – were needed to meet the requirements of human language(s) (Clark *et al.*, 2010). Thus, as linguists were focused on the detail, while advocating that computers would be of use only when they were able to see language as linguists do, computer scientists were somewhat more liberal; their aim has not been focused on having computers do what humans do when analysing language, but rather have the machine perform as well as possible, while establishing an error margin. In this sense, whereas for linguists computers are only acceptable when they get their answers 100% right, for computer scientists what is important is, not only to get the answer right – or as close as possible to 100% of the time –, but also to know how wrong the system has gone. Therefore, to the degree of detail advocated by linguists, computer scientists responded with other, more general computational devices and probability models that allowed them to increasingly provide results that, although not perfect – and especially not providing a 100% degree of reliability –, were as good as, or hopefully better than those usually provided by 'manual' linguistic analysis alone.

These systems based on probabilistic models have been at the centre of most approaches to natural language processing (NLP), and while they challenged the practice of 'traditional' linguistic analysis, they also offered linguists new and previously unthinkable possibilities. In forensic contexts, in particular, a proposal consisting of statistically gaining comprehensive knowledge of the world, in addition to knowledge of a language – as probabilistic models do – seems more appropriate than more fundamentalist proposals that argue for heavily rule-based systems learnt from scratch for processing natural language. Methodologically, one obvious advantage of probabilistic models over rule-based systems is that they build, not upon direct experience, but rather upon huge amounts of textual data produced by native speakers of (a) natural language. For applied linguists, choosing between probabilistic models and rule-based systems would be like choosing between analysing data observed by the self or analysing naturally-occurring corpus data. Another advantage is the ability to quantify the findings: as systems have been working based on statistical natural language processing (NLP) (which consists of computing, for each alternative available, a degree of probability, and accepting the most probable (Kay, 2003)), statistical models allow linguists working in forensic contexts to quantify their findings and their degree of certainty when asked by the courts. However, unlike linguists, natural language processing systems (e.g. those based on machine learning and artificial intelligence) are in general unable to indicate exactly *where* they have gone wrong, even if they are able to tell *how* wrong they are. One of the main criticisms of NLP systems is that they have so far been unable to reach the fine-grained analysis that linguists do Woolls (2010: 590), so their use in forensic contexts may be very limited, if not close to null.

Notwithstanding, as argued by Kay (2003: xx), computational linguistics can make a substantial contribution to linguistics, by offering a computational and a technological component that improves its analytic capacities. As computational systems offer linguists the ability to consistently process large quantities of text easily and quickly, while avoiding the human fatigue element (Woolls, 2010: 590), the question is not whether a perfect computational system can be designed to replace the work of the forensic linguist, but whether a simultaneous and mutual collaboration can be established between

computational and forensic linguists that provides the latter with reliable computational tools to assist their human analysis.

This article is structured as follows: the next section explains how this brief review was conducted. The subsequent sections identify some of the areas in which forensic linguists have been called upon to assist as experts, such as authorship analysis, authorship profiling and stylometry, plagiarism detection and analysis, disputed meanings, stance detection, hyperpartisanism and fake news, fraud detection, and cybercrime. Potential applications of computational linguistic systems to some of these areas are discussed, on the grounds that these are some of the applications of forensic linguistic analysis that can hardly be conducted without computational assistance. The article concludes with a discussion of some of the future challenges facing computational forensic linguistics.

Data and methodology

Research surveys are demanding methodologically, as they usually involve a systematic collection and analysis of research articles and a subsequent discussion of each individual contribution. To conduct a survey, one can either (a) perform a general search, online and in hardcopy sources, (b) focus on a keyword search in a range of reliable reference databases, (c) limit the search to a small number of benchmarking journals, or (d) select all the references published in the field within a specific timeframe. Any of these methods offers a thorough coverage across a specific period of time or range of references. However, restricting the survey to one of these approaches can be problematic in areas with an extensive range of publications, where, given the extension of the survey, the systematic analysis becomes impractical or of little use to the reader. In these cases, restricting the survey to a specific timeframe can be helpful, as it makes the survey manageable; the downside to this approach is that it limits the scope of the survey to a date interval, which doesn't necessarily mean that it is the timeframe with the most relevant publications, or when most advances have been made in the field, or the one offering the most sound basis for subsequent research.

Computational forensic linguistics is one of the areas in which conducting a survey is problematic. Firstly, given the complexity underlying the analysis of language by computers, the number of references published that address a minor language detail is enormous. An online search of the keyword 'computational forensic linguistics' in a database such as Google Scholar returned thousands of hits, and similar results are obtained in academic and scientific reference databases. Secondly, this figure increases exponentially when we consider different languages, rather than restricting the search to English. Curious readers might like to try for themselves, by searching keywords such as 'linguística forense computacional', 'lingüística forense computacional', 'linguistica forense computazionale', 'rechnerforensische Sprachwissenschaft' or others. Restricting the survey to a set date interval would not be appropriate in this area, either, since a lot of relevant research has been published over the last decades that would be left out if the survey focused on a particular timeframe.

Therefore, since not only is the number of references published over the years too extensive to allow for a systematic survey of computational linguistics methods and systems, but also highly relevant resources have been published over time, so this article focuses on a selection of references that have contributed in some way to different aspects of computational forensic linguistics. A brief survey is thus produced covering

a range of publications that I have found helpful for my own research over the years. This is accompanied by a discussion of some of the systems that can hopefully be of use to forensic linguists interested in including computational forensic linguistics in their research and practice.

Corpus Linguistics and Computational Linguistics

Applied (and, to some extent, theoretical) linguists have since the 1980s relied on corpora for research and practice. In order to make assumptions about linguistic events and language use, linguists usually rely on large volumes of spoken and/or written linguistic data that have been produced as a result of communication in context: a corpus. Although a corpus has been defined simply as "a large body of linguistic evidence typically composed of attested language use" (McEnery, 2003: 449), Bowker and Pearson (2002: 9) argue that in addition to being large and containing authentic data, a corpus needs to be available in electronic form so that it can be processed by a computer. Therefore, although a distinction is made between Corpus Linguistics and Computational Linguistics, the former can only exist as part of the latter, not only because in order to be available in electronic form, a corpus has to be subject to natural language processing, but also because some of the procedures applied to corpora (such as annotation) require sophisticated processing procedures and furthermore because corpora should ideally be tailored to be used in NLP systems. Additionally, not every set of data can be called a corpus; the collection of data needs to be well-organised (McEnery, 2003: 449) and meet some specific criteria in order to be used as a representative sample of the (subset/dialect/register/sociolect etc. of the) language that the researcher intends to study (Bowker and Pearson, 2002: 9). This will allow the linguist to make safe assumptions, while averaging out idiosyncrasies and avoiding bias. Additionally, the corpus must also take into account the time frame in which the texts were produced, depending e.g. on whether the study is synchronic or diachronic.

Given their potential to demonstrate real language use, corpora (and corpus linguistic techniques) have been widely used by forensic linguists both as part of research and in casework. As researchers and practitioners, forensic linguists can either build their own corpora or resort to ready-made corpora already available, which often operate as reference corpora. Available corpora include, among others, the BNC – British National Corpus (http://www.natcorp.ox.ac.uk), the BYU Corpora (https://corpus.byu.edu), the BYU-BNC – British National Corpus at BYU (https://corpus.byu.edu/bnc/), corpora of Portuguese (https://www.linguateca.pt/ACDC/) and the BYU Corpus de Português (https://www.corpusdoportugues.org), the BYU Corpus del Español (https://www.corpusdelespanol.org), the Corpus de Referencia del Español Actual (CREA) of the Real Academia Española (http://corpus.rae.es/creanet.html), the COMPARA – Parallel Literary Corpus (https://www.linguateca.pt/COMPARA/), parallel corpora CORTrad (https://www.linguateca.pt), the COCA – Corpus of Contemporary American English (https://corpus.byu.edu/coca/), as well as specialised language corpora, such as the Corpus of US Supreme Court Opinions (https://corpus.byu.edu/scotus/). Nevertheless, do-it-yourself (DIY) corpora (Maia, 1997) are often used by forensic linguists when conducting research or working on cases. As they have the advantage of not requiring computers with great processing capacity, and in addition can be tailor-made to suit the needs of the research project or the particular case, they allow the forensic linguist to address a particular aspect of language to which ready-made corpora may be unable to respond.

This option also offers another advantage: as DIY corpora are usually saved in the user's computer, rather than being made available in cloud systems, it provides a tighter control over the integrity of the data.

Publications on forensic linguistics that have drawn upon access to corpora – either ready-made or DIY – abound. An example of the latter is the research conducted by Finegan (2010), where the author discusses how corpus linguistics approaches can be used to analyse the adverbial expression of attitude and emphasis in legal writing, and in particular in the United States Supreme Court opinions. As, according to the author, American jurisprudence relies to a large extent on the written opinions of appellate courts, a forensic linguistic analysis of the details of legal language (in this case, adverbial expressions of attitude and emphasis) employed in those opinions can be of relevance, not only to the training offered to lawyers, but also to a deeper understanding of the legal opinions. Finegan (2010) supports his analysis of adverbial expressions of attitudinal stance and emphasis on a series of excerpts extracted from the DIY corpus of supreme court opinions (COSCO). COSCO includes a compilation of court opinions from 2008 that were not unanimous – i.e., it includes only decisions with at least one dissenting opinion, in order to simultaneously exclude procedural matters, while including differences of opinion that are more likely to reveal expressions of attitude and emphasis. The corpus contains 905,464 words overall, collected from the Lexis-Nexis database: approximately 259,000 words for opinions for California cases (17) and 647,000 words for opinions for federal cases (56), decisions that were not unanimously made by the supreme courts of California (17 cases) and by federal courts (56 cases). In order to make assumptions of the use of adverbials in supreme court opinions, Finegan (2010) calculated the frequency of stance adverbials and emphatic adverbials in COSCO and compared them against the frequency of such adverbials in general language (ready-made) corpora, namely the BNC and the BROWN corpus, to conclude that their use in supreme court decisions is more frequent than in general language. Based on this study, the author discusses the efficacy of emphatics in appellate briefs, and especially wonders whether using those adverbials found comes as a disadvantage. Finegan (2010) thus shows how (the computational processing of) corpora can be used to fully and accurately describe legal language, which, as he advocates, is a responsibility of forensic linguists.

Corpus linguistics, and its underlying computational approaches, has also been used to conduct research into forensic authorship analysis. It is generally accepted that one of the assumptions of forensic authorship analysis is the existence of idiolect, i.e. "the theoretical position that every native speaker has their own distinct and individual version of the language they speak and write, their own idiolect" (Coulthard, 2004: 31), even if the difficulty in empirically substantiating a theory of idiolect has given rise to concerns that the concept itself is too abstract to be of practical use (Grant, 2010; Turell, 2010). Empirically-driven research, however, exists. In their study, Johnson and Wright (2014) discuss how stylistic, corpus, and computational approaches to text have the potential to identify *n*-grams, and be used for authorship attribution in a way that is similar to the one that journalists use to identify relevant soundbites. These the authors call '*n*-gram textbites' (Johnson and Wright, 2014: 38). In order to investigate whether 'n-gram textbites' are characteristic of an author's writing, and whether those chunks of text can operate as DNA-like identifying material, the authors conduct a case study based on the computational analysis of the Enron corpus. This corpus includes 63,000 emails (totalling

2.5 million words) written by 176 employees of the former American energy corporation Enron. The analysis of the n-grams extracted from the corpus, and the subsequent stylistic analysis, reveals that one Enron employee uses politely encoded directives repeatedly, thus building a habitual stylistic pattern. A statistical experiment conducted with anonymised texts of the same author demonstrated that the use of word n-grams as 'textbites' could successfully attribute larger samples of text to the same author, while even smaller samples reported promising results.

Authorship analysis, authorship profiling and stylometry

Authorship analysis, and especially stylometric approaches to authorship analysis, has been one of the forensic linguistic applications that has probably attracted most of the interest of computer scientists working in natural language processing. As a simple web search demonstrates, the question 'who wrote this text?' has long intrigued computer scientists, who have dedicated time and effort to investigate the authorship of literary and non-literary texts alike. In some cases, software packages were developed based on the research conducted; an example is the stylometric analysis software Signature[1], which is largely based on the analysis of 'The Federalist Papers'. Over time, however, as computers gave answers to the less complex questions, new challenges were taken on-board, and the degree of sophistication of the questions increased.

One example of these challenges is described in the research conducted by Sarwar *et al.* (2018), who approach the topic of cross-lingual authorship identification. Given labelled documents written by an author in one language, the authors aim to identify the author of an anonymous document written in another language. One of the main challenges of cross-lingual authorship identification is that, as is well known to forensic linguists, stylistic markers vary significantly across languages. To overcome this problem, it is reported that methods such as machine translation and part-of-speech tagging can be useful, except when dealing with languages for which such resources are nonexistent. This, together with the fact that, as the authors state, the performance of such methods tends to decrease as the number of candidate authors and/or the number of languages in the corpus increases, brings additional challenges for use in forensic linguistic contexts. In order to overcome these issues and enable cross-lingual authorship identification, the authors analyse different types of stylometric features and identify 10 features that they claim are language-independent, and furthermore are of high performance. These features include measures of vocabulary richness, structural features (average number of words per sentence and number of sentences in the chunk), and punctuation frequencies (frequency of quotations, frequency of punctuation, frequency of commas, and frequency of special characters). The method adopted, which consists of partitioning the documents into fragments and then decomposing each fragment into fixed size chunks (of 30,000 tokens each), is reported to yield a very good level of accuracy: 96.66%, using a multilingual corpus of 400 authors with 825 documents written in 6 different languages. Impressive as this may be, however, sample size is a crucial issue in forensic contexts: although forensic linguists are sometimes given access to considerably high volumes of text, large samples are rare and in most cases linguists have to cope with small samples, in which case the system might be less efficient.

Amelin *et al.* (2018) also report on their work on the analysis of the dynamic similarity of different authors to identify patterns in the evolution of their writing style. One of the main shortcomings of this study is that the method has been tried and tested with

literary works, and not with text that has been produced spontaneously, and even less so with forensic texts. Therefore, it can be hard to tell whether changes in patterns derive from the evolution of the authors' writing style, or are features of the literary persona, or due to literary edits, by one or more editors – i.e., multi-authored texts. Notwithstanding, the method could have some merit if applied to forensic contexts, as it could potentially be useful to establish intra-author variation. Stylometry has also been of huge interest to computational linguists, not only as an approach to identify the style of an author of literary works, but also in an attempt to attribute the authorship of suspect or unknown texts. That is, for example, the case of Neme *et al.* (2015), who employ algorithms to identify stylistic attributes (and resolve anomalies), allocate a set to one of several possible classes (classification) and offer a visualisation structure. The visualisation system, in particular, could be of interest to forensic linguists, but again the method remains on the literary level, as it is not applied to non-literary texts, and even less so to forensic texts.

A more forensic-grounded research is presented by Paul *et al.* (2018), who address the issue of divergent editorial identities resulting from freedom of editing, and which often negatively impact the integrity of the data – and consequently of the editorial process – in the form of malicious edits and vandalisation, among others. The authors argue that malicious behaviour of ambiguous identities can be resolved, at least in part, by disambiguating the users' identity, which allows a distinction between trusted and mischievous users. However, unlike other studies that they report in the literature, the method that they propose does not use linguistic features for authorship analysis.

In the same vein, Zhang *et al.* (2014) state that, in addition to literary works, the authorship identification of authors of anonymous texts is particularly relevant in areas like intelligence, criminal law, civil law and computer forensics. The authors thus propose a semantic association model that takes into account voice (the relationship between a verb and the subject of the action), word dependency relations, and non-subject stylistic words (words that are not related to the topic of the texts) to enable a representation of the writing style of unstructured texts of various authors. Subsequently, an unsupervised approach is designed to extract stylistic features, and employ principal component analysis and linear discriminant analysis to identify the authorship of the texts. Although the authors report that, by capturing syntactic and semantic stylistic characteristics involving words and phrases, this approach significantly improves the overall performance of authorship identification, they also admit to the existence of some challenges and difficulties to computational authorship identification, such as the number of candidate authors, the size of each text, and the number and types of training texts, in addition to issues related to language, genre, topic, stylistic features and available documents. Such difficulties, as the authors agree, make it difficult for computers to extract the stylistic characteristics of different types of texts, and establish the authorship of those texts. The authors recognise that this is especially difficult in forensic cases, where the quantity – and size – of the texts available for investigation, as previously mentioned, is usually small.

A range of the references surveyed show that computational forensic linguistics has been largely dominated by computer scientists with an interest in linguistics. Although good to excellent results have been achieved by many of these systems, the interest of computer scientists lies mainly with the capacity of the machine to process information

and achieve the best possible results – while establishing the *precision* (percentage of texts correctly attributed to an author among all the texts attributed) , *recall* (percentage of texts written by an author that were attributed correctly over the total number of texts written by that author) and F_1 (average of precision and recall) –, more than it does on making safe assumptions for investigative and mainly evidential purposes. Conversely, linguistics studies that resort to computer science to support their analysis are less common, although they exist. In the field of authorship analysis, Nini (2018) conducted an authorship clustering/verification analysis of the letters purportedly written by Jack the Ripper in order to investigate whether a different author may have written the earliest texts, as some theories argue that these texts were written by journalists with the aim of selling more newspapers. A cluster analysis of the corpus of 209 letters was conducted using the *Jaccard* distance of word bigrams. The quantitative analysis conducted, together with the identification of some shared distinctive lexicogrammatical structures, led the author to conclude that these findings support the hypothesis that, not only were the two most historically important letters written by the same person, but also there is a link between these two texts and the *Moab and Midian* letter, which is another key text in the case.

More recently, Grieve *et al.* (2018) discuss the use of computational forensic linguistics in the famous case of the 'Bixby Letter'. The 'Bixby Letter' is a letter of condolence that was sent by the late President of the USA Abraham Lincoln to Lydia Bixby, a widow that was believed to have lost several sons in the Civil War. The letter is considered a remarkable piece of correspondence, in no small part due to the writing style of the author. However, the authorship of the letter has not been unquestioned. Although the letter was signed by Lincoln, some historians argue that its true author was John Hay, who was then Lincoln's personal assistant. One of the difficulties in attributing the authorship of the letter is its length: as the letter is only 139 words long, standard techniques are ineffective, which largely accounts for disappointing previous authorship analyses, which have been inconclusive. Grieve *et al.* (2018) point three issues when manually selecting the linguistic features for analysis, especially in cases of short texts: (1) the selection of the most relevant linguistic features depends on the analyst, which helps to explain the lack of agreement among analysts; (2) the variation in the amount of material available as writing samples of the possible authors is difficult to control; (3) the differences reported in the usage of the linguistic forms are difficult to judge, as it is difficult to determine whether they are sufficient to attribute authorship reliably. (The findings of the authors are discussed below.)

Indeed, sample size is one of the most relevant methodological challenges to authorship analysis. Although forensic linguists constantly have to analyse short texts in forensic contexts (Coulthard, 2004; Coulthard *et al.*, 2017), such texts raise particular methodological issues, as they cannot usually be analysed using quantitative, statistical methods. Unsurprisingly, therefore, Stamatatos (2009a: 553) called it 'the most important' methodological issue in the area. This issue has been the focus of research into forensic authorship analysis for some time. Yet, previous computational studies have shown some promising results with small text samples. For example, research previously conducted on the authorship attribution of Twitter messages demonstrated that short messages can be successfully and accurately attributed computationally (Sousa-Silva *et al.*, 2011). This research focused on an aggregate set of features, including quantitative markers (e.g.

text statistics), markers of emotion (e.g. smileys, 'LOLs', and interjections), punctuation and abbreviations. Support Vector Machines (SVM) were used as the classification algorithm, given their robustness, using a *1-vs-all* classification strategy. For each author, a SVM was used to learn the corresponding stylistic model, so as to be able to discriminate each author's messages. The method, which combined text classification techniques and a group of content-agnostic features, reported very good results in successfully attributing the authorship of Twitter messages to three different authors. This study was innovative in that automatic authorship attribution of text strings as short as the ones described (i.e., up to 140 characters) using only content-agnostic stylistic features had not been addressed before. The study showed that a relatively small volume of training data (i.e., texts of known authorship) is required; as little as 100 messages of known authorship are sufficient to achieve a good performance in discriminating authorship.

In the study conducted by Grieve *et al.* (2018), the authors propose a method to which they call *n-gram tracing*, which combines stylometric and forensic stylistic analysis, to conduct a quantitative analysis of short text messages. The method consists of extracting sequences of character and word *n*-grams in the questioned document and calculating the percentage of all *n*-grams occurring at least once in each corpus and finding the author with the higher percentage of those forms – or with the larger number of unique *n*-grams. One of the benefits of the method, the authors argue, is that it allows an extraction of all possible features in each corpus; the other is that it considers the existence or absence of the different features, rather than their relative frequencies. In other words, the method proposed consists of measuring the set of *n*-grams found in the questioned document and in each set of documents of each possible author. The questioned document "is then attributed to the possible author with the highest overlap coefficient" (Grieve *et al.*, 2018: 7).

Although the general applicability of the *n*-gram tracing method is neither assessed, nor assumed in the research conducted, the authors cite Grant (2013) to argue that this is not a prerequisite to apply a method in a particular forensic authorship analysis case. Notwithstanding, the authors measure the accuracy of the method, namely the *precision* and *recall* scores, as well as the F_1 score. The findings report F_1 scores in the analysis of character *n*-grams of at least 0.95 for both authors on analyses between 5-10 characters, with the best results obtained at 7-8 characters. The authors also report excellent results when attributing authorship based on at least 4 of the 7 analyses: the author of all 1,662 texts was correctly identified. Similarly, good results were obtained when computing word *n*-grams: the authors report F_1 scores above 0.90 on analyses of unigrams to trigrams for both authors, although bigrams are the best performers, with F_1 scores of 0.96 for Lincoln and 0.94 for Hay. As reported by the authors, the analyses of 4- to 16-character *n*-grams and 1- to 3-word *n*-grams were particularly useful for distinguishing between the writings of Lincoln and Hay. Based on these findings, the authors conclude that the sequences that perform better are those that are neither too short (and that consequently tend to be reused by all authors), nor too long (and consequently tend to be used by none of the authors). They also argue that selecting features manually can be misleading, particularly when those features are rare. The authors therefore propose a simple method that is based on extracting all the features of a particular type occurring within a text.

Plagiarism detection and analysis

A controversial issue in computational plagiarism detection is its own definition. As previously stated (Sousa-Silva, 2013), the concept of plagiarism is too complex to allow computers to detect it. Some commercial systems, for example, are unable to identify a word as having been plagiarised simply if changes in spelling (resulting from writing in different language variants) are introduced. Therefore, as then argued, at most computer systems are able to detect textual overlap. Notwithstanding, a simple web search using the search phrase 'plagiarism detection' is indicative of how commercial systems market themselves.

Plagiarism detection remains one of the main areas of research in the field of computational linguistics, and the field has long attracted interest from research and industry organisations (Potthast *et al.*, 2009). This is unsurprising, if one takes into account that: (a) commercial plagiarism detection systems have been developed worldwide, in order to assist teaching staff, (higher) education institutions and publishers, among others, with the identification of improper text reuse – while, of course, retaining their focus on profit margins; (b) plagiarism strategies and techniques have evolved over time, and so has the technology used, so new methods and approaches are required to detect plagiarism – consequently, permanent research is necessary to keep systems up to date to address those challenges.

Nevertheless, many challenges remain to computational plagiarism detection, the most basic of which is probably the fact that computers can only detect textual overlap, but not whether it is as a result of plagiarism. Indeed, in academic and non-academic contexts alike, textual overlap does not necessarily equate with plagiarism, and real cases abound of instances of textual overlap that are not plagiarism. This is a crucial distinction, which should lie at the basis of any plagiarism detection approach, as simply terming computational systems that identify textual overlap 'plagiarism detection software' is misleading; in order to judge an instance of textual overlap as plagiarism, a detailed linguistic analysis is required that considers, e.g., the amount of textual overlap, use of unique vocabulary and/or phrases, volume of verbatim copying vs. text edits, use of paraphrasing and rephrasing, strategies of coherence and cohesion, and translation, not to mention prior authorship. Therefore, simply assuming that there is a plagiarism threshold, and consequently that a lower or higher volume of textual overlap is synonymous with the absence or existence of plagiarism, can bring along serious risks of falsely making or otherwise discarding plagiarism accusations.

In forensic contexts, linguistics-focused computational systems have demonstrated greater reliability than purely computational, statistics-based models. Woolls and Coulthard (1998), for example, show how two computational tools that were not initially designed for forensic linguistic analysis demonstrated being extremely useful for plagiarism detection: *Toolkit Analyser* and *FileComp*. Among other specificities, the former allowed forensic linguistics to calculate lexical richness quickly and easily, while the latter was designed to allow users to compare and contrast two or three files against each other and produce details about shared and unique vocabulary (both of which are crucial in analysing plagiarism). The usefulness of the system and its successor *CopyCatch* (Woolls, 2003) was demonstrated by Johnson (1997) and later by Turell (2004) in academic and forensic cases. In particular, the fact that this software allows a comparison of lexical items across different texts, after removing stop words, allows forensic

linguists to analyse instances of potential plagiarism, regardless of the order in which the words are presented in the original and in the suspect texts.

Research into computational plagiarism detection has continued in all directions, however, which eventually enabled the identification of plagiarism patterns that were previously unthinkable. In general, computational plagiarism detection has focused on information retrieval, a computer science task that consists of searching for information in a document, or searching for documents themselves. The research conducted within the scope of the PAN competition is an illustrative example in this respect. PAN is 'a series of scientific events and shared tasks on digital text forensics and stylometry' (https://pan.webis.de/), whose competitions have been running since 2009, when the first International Competition on Plagiarism Detection took place. Although the data-sets that have been used over the years do not necessarily consist of forensic texts, they can still give some insight into possible approaches to forensic problems. The first competition, for example, aimed to establish an evaluation framework for plagiarism detection systems (Potthast *et al.*, 2009), by providing a large plagiarism corpus against which the quality of plagiarism detection systems could be measured. This evaluation framework consisted of four phases: an external plagiarism detection task, an intrinsic plagiarism detection task, a training phase and a competition phase. As the authors argue, one of the reasons why such "a standardized evaluation framework" (Clough, 2003) is nonexistent is that even commercial plagiarism detection systems were unavailable for scrutiny – and so they remain.

In the PAN competition, plagiarism detection was divided into 'external plagiarism detection' and 'intrinsic plagiarism detection'; the first is used to refer to a case where a suspect text is compared against the potential (expected) originals (Stein *et al.*, 2007), whereas the latter is used to refer to a case where a text is suspected to be plagiarism, but no sources are available against which to compare it (Meyer Zu Eissen and Stein, 2006). In this latter case, the text is analysed intrinsically; the analysis thus focuses on trying to identify relevant stylistic cues that may be indicative of shifts in the writing style of the author. In these cases, the suspicion is raised, not intuitively (as happens when a lecturer notices shifts in style while marking a student's essay), but computationally, by resorting to a stylistic analysis. The intrinsic plagiarism detection approach can be extremely useful, especially as the potential sources are not available for comparison, despite some of its shortcomings, from a forensic linguistics perspective, which are related to the circumstances of the academic text genre, and which will be discussed below.

The plagiarism corpus provided for the PAN competition consists of texts written in English, and includes 41,223 texts with 94,202 cases of automatically inserted plagiarism. The instances of plagiarism inserted in the corpus range between 50 and 5,000 words and include same-language plagiarism, as well 10% of text that was lifted from text excerpts written originally in German and Spanish, and then machine-translated into English. The corpus also includes some instances of obfuscation 'random text operations' (such as shuffling, removing, inserting, or replacing words or short phrases at random), 'semantic word variation' (i.e., randomly replacing lexical items with synonyms, antonyms, hyponyms and hypernyms) and 'POS-preserving Word Shuffling' (in which words in the sentence are shuffled, while retaining the POS (parts-of-speech) order) (Potthast *et al.*, 2009).

In the first competition, 10 (out of 13) systems were submitted for the external plagiarism detection task and 4 were submitted for the intrinsic plagiarism detection task. In the case of external plagiarism, only 6 systems showed a noteworthy performance, with the system described by Grozea *et al.* (2009) winning the competition. This system is based on establishing a similarity value based on n-grams between each source and each suspicious document, and then investigating each suspect pair in more detail in order to determine the position and length of the texts that have been lifted. One of the most striking features of this system is its processing capacity: in 2009, a single computer was able to compare more than 49 million document pairs in 12 hours. In the case of intrinsic plagiarism detection, only one system performed above the baseline: that of Stamatatos (2009b). In this system, the author uses character *n*-gram profiles and a function to identify style changes that builds upon dissimilarity measurements in order to quantify style variation within a given document. This method is based on the system originally proposed for author identification (Stamatatos, 2006). Although each system was the best performer in each task (and hence winners of the competition given their good performance), the rates of precision and recall in both cases are far from those expected from forensic linguists, as precision scores of 0.74 and 0.23, in the external and intrinsic plagiarism detection tasks, respectively, are not sufficiently good for forensic scenarios. Subsequent PAN competitions (namely, the second competition, in 2010, and the third competition, in 2011) revealed some improvement in the precision, recall and granularity rates (against which the systems' performance has been measured), but not significantly. For example, in the second competition (2010), in which the external and intrinsic plagiarism detection tasks were combined in one single task, the winning system (Kasprzak and Brandejs, 2010) showed a recall of 0.6915 and a precision of 0.9405 when tested over the external plagiarism data alone. In the 2011 competition, all the top three plagiarism detectors built upon the results obtained by systems submitted in previous years (Potthast *et al.*, 2011): Grman and Ravas (2011), Grozea and Popescu (2011) and Oberreuter *et al.* (2011).

For forensic linguists, the methodology used in this competition can raise some important issues. The first is that, in contexts like the academic, not only are writers allowed to integrate other people's voices in their own text, they are also expected to do so. Also, especially in cases of 'patchwriting' (Howard, 1995), where students are in the process of learning how to write academically by resorting to the sources, an inconsistent writing style is to be expected. Therefore, 'blindly' relying on the computational analysis may – again – give rise to false positives. In other words, those systems are unable to account for – and discount – instances of text legitimately quoted from other sources, they do not account for different authorial stances that are merged in the text, and perhaps even more importantly, they do not take into account the fact that the writer may still be learning how to write academically. Therefore, as Potthast *et al.* (2009) aptly point out, that kind of analysis requires human analysts to make grounded decisions as to whether it is a case of plagiarism or not. An additional issue for forensic linguistic applications is that the method has been tested a corpus of artificially-created plagiarism, and not on a corpus of naturally-occurring plagiarism. While forensic linguists usually find it acceptable to train and experiment with non-forensic data, when such data are unavailable, it is a requirement that the data are at least naturally-occurring. Interestingly, however, plagiarism is inherently a creative task, which consists of constantly inventing

new ways to deceive – so, in this respect, the methods underlying the PAN corpus are to some extent realistic. In any case, the worth of the system as a computer-assisted plagiarism detection tool is undeniable.

Abdi *et al.* (2017) critique the most commonly-used approach to plagiarism detection, which consists of comparing the surface text of a suspect document against that of a given source document, on the grounds that alterations introduced to the text (such as changing actives to passives and vice-versa, changing the word order, or rephrasing the text) may interfere with the plagiarism detection, and offer misleading results – either by producing false negatives (thus missing actual instances of plagiarism) or false positives (resulting e.g. from strings of text that are commonly used and not necessarily unique). The method proposed by the authors (IEPDM) to overcome these issues consists of using syntactic information (namely, word order), content word expansion and Semantic Role Labelling (SRL). The task of SRL is to analyse a sentence, starting with the verbs, in order to recognise all the constituents that fill a semantic role (Carreras and Màrquez, 2005). The aim of the content word expansion approach is to enable the identification of similar ideas expressed using different words Abdi *et al.* (2017). Overall, the authors report that the method proposed is able to detect different types of plagiarism, from verbatim copying to paraphrasing, including changes to sentences and word order, and overall perform better than existing techniques and better than the four top-performing systems competing in PAN-PC-11. Nevertheless, although the results reported are very good when compared to other systems (*plagdet* score of 0.735, when compared to the PAN-PC-11 *plagdet* score of of 0.675), and any computational approach that helps the human analyst identify potential cases of plagiarism, the system is still far from ideal for accurate plagiarism detection in forensic cases.

Conversely, Vani and Gupta (2017) propose a binary approach to plagiarism detection based on classification using syntactic features, as a means to establish whether a suspect text *is* – or conversely *is not* – an instance of plagiarism. The authors extract linguistic features based on syntax, by applying shallow natural language processing techniques – i.e., part-of-speech (POS) tags and chunks – to propose this method as an intermediate analysis, before running exhaustive and detailed analyses of the text passages. This method has great potential in establishing whether a document is likely to have been plagiarised, before asking the analyst to make a decision as to whether the suspect text needs to be analysed further, by subsequently running careful and detailed analytical procedures, which are usually time-consuming. This research is explored further (Vani and Gupta, 2018), by combining a syntactic-semantic similarity metric taking into account POS tags, chunks and semantic roles; the latter built on the extraction of semantic concepts from the WordNet lexical database. To test this method, the authors resort to the corpus released yearly by the PAN competition between 2009 and 2014, and report a performance that is better than the top-ranked performers of each year. In the case of the former study, the authors conclude that the fact that the results obtained outperform the baseline approaches demonstrates the convincingness of using syntactic linguistic features in document level plagiarism classification; yet, although reference is made to instances that are close to manual or real plagiarism scenarios, the extent to which the methods work with real, forensic cases of plagiarism is unknown.

One area in which plagiarism detection and analysis is increasingly relevant is journal editing. Over the last decades, not only has the number of journals increased expo-

nentially, but also the number of 'predatory journals' has significantly increased. This, on the one hand, encouraged the multiplication of identical submissions by author(s) as a result of the pressure put on researchers to publish, while, on the other, encouraging the submission of replicated, plagiarising material in those predatory publications. In order to assist them in making informed decisions on whether to publish, publishers and journal editors alike would greatly benefit from computer systems that allow them to identify potentially unoriginal material quickly and efficiently.

The method proposed by HaCohen-Kerner and Tayeb (2017) goes in this direction: a two-stage process is suggested, which consists of (1) filtering the suspect and non-suspect text, in order to discard those that fall below the 20% threshold, and (2) applying 3 novel fingerprinting methods to the suspect texts – i.e., those texts whose similarity with other sources is equal to or higher than the threshold. Traditionally, fingerprinting techniques have used character *n*-grams (Butakov and Scherbinin, 2009), word *n*-grams (Hoad and Zobel, 2003), sentences (Barrón-Cedeño and Rosso, 2009), or a combination of different methods (Sorokina *et al.*, 2006) to identify document similarity. HaCohen-Kerner and Tayeb (2017) use a combination of three prototype fingerprinting methods to compare the tested papers and the retrieved papers across three dimensions, and thus establish the extent of document similarity. The authors report an improvement, as compared to previous heuristic methods.

As previously discussed (Sousa-Silva, 2013), it has long been established that some instances of plagiarism can hardly be detected without human investigation (Maurer *et al.*, 2006; Mozgovoy, 2008). Among the set of limitations imposed on plagiarism detection systems is the most important of all: the inability to detect plagiarism; at most, the so-called plagiarism detection systems can establish the degree of similarity between documents, and produce some scores to report the amount of potentially overlapping text. Obviously, the availability of a system that produces such scores can be, in itself, of great help to the human analyst, who can start the forensic linguistic analysis with the machine-calculated similarity scores and then move on to establish whether it is a case of plagiarism. Among the biggest challenges for machine plagiarism detection, Maurer *et al.* (2006) pointed to (1) the use of paraphrasing, (2) the unavailability of comparison documents in electronic form, and (3) translation. They predicted that there was hope for challenge (2), since the world is becoming increasingly digitised; (1) is the one for which most progress would be expected, given the technological developments in paraphrasing analysis and detection; (3), on the contrary, would remain a challenge for some time. Research conducted in subsequent years, however, demonstrated that some of the authors' predictions failed, since as discussed in Sousa-Silva (2013) and Sousa-Silva (2014), plagiarism by translation – i.e. where translation is used to pass off someone else's text, work or ideas as one's own – can now be effectivelly detected, whereas detecting plagiarism resulting from e.g. the use of paraphrasing strategies remains a challenge.

In their work, Barrón-Cedeño *et al.* (2013) address the issue of translated plagiarism (which they call 'cross-language plagiarism detection') by testing three different models to estimate cross-language similarity: (1) Cross-Language Alignment-based Similarity Analysis (CL-ASA), (2) Cross-Language Character n-Grams (CL-CNG), and (3) Translation plus Monolingual Analysis (T + MA). (1) uses a computational algorithm to establish the likelihood that a suspect text has been translated from a text in another language; (2) consists of removing all punctuation, diacritics and line breaks, among others, to struc-

ture the text into character *n*-grams to estimate the similarity between two documents; (3) consists of translating all documents into one common language (English), removing stop-words, lemmatising them, and then comparing the texts. The model described in (3) obtained the best results, with an F_1 score of 0.36 – when compared to F_1 scores of 0.31 and 0.15 of models (1) and (2) respectively. The potential of the system relies on the fact that, as Barrón-Cedeño *et al.* (2013) claim, if the system marks a text as suspect, then that text is worth being investigated further by a human; however, it is still far from the fine-grain required by forensic linguists to analyse and detect plagiarism.

A different approach is adopted by Pataki (2012), who describes a method for translation-based plagiarism based on establishing the distance between sentences, which are subsequently evaluated in multiple steps. The aim is that the system allows a comparison of all possible translations, rather than giving precedence to a translation offered by a machine-translation system. The author uses the Hungarian-English language pair, but claims that the system is robust with any pair of European languages. This system operates based on three steps: (1) a search space reduction is performed; the text is split into smaller chunks (in this case, sentences), the lemmas in the chunks are identified, a bag of words containing all the translations of the lemmas is created, and stop words are removed; (2) text similarity is evaluated, using a similarity metric, previously using dictionaries; and (3) post-processing of the texts, which selects the most likely candidates. Overall, the author reports some encouraging results, although it is also admitted that there is room for improvement, as the precision scores obtained by the system did not produce relevant output. In addition, this information retrieval system was tested using an artificial test corpus. Encouraging as the results reported may be, they are very far from the those needed by forensic linguists when handling forensic plagiarism cases. Moreover, given the degree of computational sophistication and the number and the nature of resources needed, the system's usefulness in forensic contexts is disputable.

Another computationally sophisticated system to detect translation-based plagiarism is the one described by Franco-Salvador *et al.* (2016). In their study, the authors aim to investigate whether a mixed-methods approach that combines knowledge graph representations (which are generated from multilingual semantic networks) and continuous space representations (which are inherently semantic models) can contribute to improving the performance of existing methods. In this system, the estimation of the similarity between text fragments is based on an analysis of the similarity of word alignments. Tests are run by the authors in order to assess the performance of the model proposed against other existing models in detecting instances of plagiarism of different lengths and using different obfuscation techniques. These tests are performed using the PAN 2011 competition corpus (PAN-PC-2011) data-sets, which consist of texts in two language pairs: Spanish-English and German-English. The authors conclude that a method combining knowledge graphs and continuous models outperforms the results obtained by each system individually – on the grounds that, as each model captures different aspects of text, they complement each other.

The hybrid model proposed by Franco-Salvador *et al.* (2016) shows an excellent performance, especially if one takes into account that hybrid models do not always perform better than their component models individually. In addition, the authors also report an equally excellent performance in handling different types of plagiarism – including

short, medium and long instances of plagiarism, instances of machine-translated plagiarism, and instances of machine-translated plagiarism that are subsequently obfuscated manually. Notwithstanding the promising results described, this system may show some shortcomings in forensic contexts. Firstly, the data-sets used to run the tests have been artificially created, so whether using the model to analyse authentic forensic data would produce identical results is unknown. Secondly, the PAN data-sets contain very large volumes of data, especially when compared to the volume of suspect text in real, forensic cases of plagiarism; although, as the authors claim, the model is a high performer even detecting plagiarism in short excerpts, it is likely that such high performance is negatively impacted by lower volumes of text. Finally, notwithstanding the excellent results obtained, the model is likely to be of limited usefulness in forensic linguistics contexts, for reasons identical to the ones pointed out for the model described by Pataki (2012) – i.e., high level of sophistication and additional underlying resources needed.

Conversely, in forensic contexts the most commonly used methods are undoubtedly those that use existing tools and resources, rather than attempting to develop new tools. One of these methods for detecting translation-based plagiarism – or *translingual plagiarism*, as it has been termed – is the one described in Sousa-Silva (2013, 2014). The method proposed consists, firstly, of conducting a linguistic analysis of the suspect text(s) in order to identify linguistic clues that function as indices of the language of the potentially original text. The suspect text is then translated into that language using one of the several machine translation engines available (e.g. Bing, Google Translate, etc.). Next, function words are selected as stop words, while retaining lexical items; this is built on the assumption that machine translation engines usually have problems handling function words, such as prepositions and determiners, but tend to perform well when translating lexical items. Some lexical items are then selected as keywords in order to conduct an Internet search using any common search engine. Examples from previous authentic cases of plagiarism show that the method performs well in identifying the source, although it is also possible that no original texts can be found (Sousa-Silva, 2013, 2014). In the latter case, this can mean either that (a) the original is available in a language other than the one into which the suspect text was machine-translated, or (b) the suspect text is indeed original.

Although this method has been proven to work well overall, it has some drawbacks. Its shortcomings include the fact that this procedure is mostly machine-assisted, rather than automated; if APIs (Application Programming Interfaces) were available – as was once the case with Google Translate – systems could have them built in and automate some of the steps. Access to some of these APIs has, however, been revoked meanwhile, so several steps have to be performed manually by the analyst. Likewise, many of the decisions have to be made by the analyst, as is the identification of the possible language of the original. Conversely, the method offers many advantages, especially for forensic linguists. Firstly, the lower degree of automation, while requiring a stronger user intervention, offers the analyst a tighter control over the analysis. Secondly, the procedure builds upon two commonly used resources – machine-translation and search engines – that are permanently updated, without any action required from the analyst (unlike most or all of the systems previously described); this means that the analyst is able to use them freely and at any time. Finally, the method can be easily explained, justified,

and – if necessary – replicated, which is crucial in some forensic cases, especially cases that end up in court.

The future of forensic linguistics (and) computing

One of the main foci in police-related research is predictive policing, which consists of using mathematical and statistical data for purposes of predicting crimes, offenders, victims of crime and perpetrators' identities. Indeed, being able to predict and deter crime by, for example, detecting fraud, deceptive language and lies, is the 'holy grail' of policing – and forensic linguistics –, and therefore is unsurprisingly of utmost interest, both to police forces and to forensic linguists. The former, in particular, would certainly welcome a system that can help them detect deceptive language, while leaving the interviewer free to concentrate on the interviewing process. Quijano-Sánchez *et al.* (2018) discuss the relevance of using natural language processing (NLP) and machine learning (ML) techniques for forensic purposes. The authors use a data-set of more than 1,000 false cases of robbery reported to the police in 2015 to develop a system (*VeriPol*) that, upon the automated analysis of a text, helps the police officers discriminate between true and false reports. The classification model builds upon the extraction of patterns and insights used when successfully lying to the police. These patterns are distributed across four categories of variables: a binary variable; a frequency variable; a logarithm variable; and a ratio variable. The authors report that the system shows a success rate of over 91% in discriminating between true and false reports, performing 15% better than the officers, and they conclude by arguing that there is a correlation between the number of details and true reporting, so the more details, the less likely that a report is false.

Predictive policing methods, however, have been criticised in recent years for several reasons. One of the arguments against them is that purely mathematical and statistical analysis not only does not guarantee being accurate at all times, but also the results are often not statistically significant (Saunders *et al.*, 2016). Another is related to the quality of the data used in the training data-sets: Lum and Isaac (2016) examined the consequences of using biased data-sets to train such systems, and the American Civil Liberties Union issued a joint statement showing their concern and criticism of the tendency of predictive policing to encourage racial profiling (American Civil Liberties Union, 2016). Much of this criticism can potentially be addressed by complementing traditionally predictive policing methods with forensic linguistic data and approaches. The research of Grant and MacLeod (2018) is a very good example of such an approach. The authors propose a model for understanding the relationship between language and identity that, despite being primarily aimed at assisting forensic linguists in training officers in identity assumption tasks, has the potential to be used in predictive policing.

Another area where computational linguistics has made significant progress, and which can be highly relevant in forensic contexts, is fake news and hyperpartisan news detection, which are two excellent examples of illicit behaviour online, and in some cases they can even be considered cybercriminal activities, alongside other online technology-enabled crimes, including intellectual property infringement, hate speech, cyberbullying, cyberstalking, insult and defamation. Although fake news and hyperpartisan news are distinct phenomena, the two can often be intertwined, as for instance detecting radical stances for or against a certain political view can be helpful in detecting potential fake news, too. The study conducted by Allcott and Gentzkow (2017), for example, relates the two by reporting that readers tend to believe in fake news mostly when the

news is in favour of their favourite candidate/policy/topic, etc. Interestingly, forensic linguistic analysis has a very important role to play in this area, especially since it is now clear that fact-checking is far from sufficient to deter the proliferation of fake news online. An effective detection of hyperpartisan news thus has a significant potential, especially if it includes linguistic information. The study of Cruz *et al.* (forthcoming), conducted as part of the Hyperpartisan News Detection competition organised by PAN @ SemEval 2019[2], shows some promising results: the model computes some text statistics traditionally used in forensic authorship analysis that are demonstrated to be effective. As these activities, like the other cybercriminal activities mentioned, share the fact that they use language, to a lesser or greater extent, they are particularly suitable for forensic linguistic analysis.

One of the main challenges of cybercrime is user anonymity, whether real or perceived. As users feel that they remain anonymous behind the keyboard – either by creating fake user profiles, or simply hiding their identity – they tend to do and say things that they would otherwise refrain from doing in face-to-face contexts. Furthermore, that anonymity is often guaranteed by using stealth technologies, IP address hiding software, the dark web – or even simply using free access, publicly available computers such as those found in public libraries and cybercafes. In these cases, forensic authorship analysis is crucial to the investigation, as it has the potential to attribute the authorship of the questioned text(s) to a suspect. Previous case work in the field of cybercrime where forensic authorship has been successfully used include a case of intellectual property infringement (using a website and Facebook), a case of defamation (using email) and a case of cyberstalking (using mobile phone text messaging). In these cases, forensic authorship analyses have been conducted in order to establish whether the cybercriminal communications had been likely produced by the suspect(s). Other instances of cybercrime can benefit from other applications of forensic linguistic analysis, however, as is the case of hate speech and offensive language. In this case, research such as the one conducted by Butters (2012) and Shuy (2008) show some of the methodological approaches adopted by forensic linguists, and the study described by Finegan (2010) further demonstrates how to approach the problem computationally. Since the language of insult often originates a conflict of interpretations, both a linguistic-juridical and a computational forensic linguistics approach to the problem are required to inform the trier of fact as accurately as possible. Although machine learning methods and techniques can potentially be used in cases of suspect communications to help detect (suspect) meanings, ultimately a forensic linguistic analysis is essential to establish the meanings involved.

Conclusion

Computational linguistics has evolved significantly over the last decades. Increasing computer processing power, together with the growing attention of computer scientists to natural language processing (NLP), has enabled more in-depth research into computational and computer-assisted linguistic analysis. Sophisticated computational systems and models have been developed that allow an analysis of large volumes of linguistic data with little human intervention, at a pace and with a degree of efficiency against which linguists can hardly compete. Interestingly, until recently linguists have demonstrated a comparatively smaller interest in computers than computer scientists in language. This is clearly shown by the research surveyed in this article, most of which has been conducted by computer scientists. It is a fact that computational linguistics should

ideally be handled by interdisciplinary teams of linguists and computer scientists. However, this does not mean that linguists cannot be – or cannot act as – computational linguists, rather on the contrary; even if linguists fall short of the advanced programming skills of computer scientists, they have the knowledge required to (a) assess the worth of computational resources under specific circumstances, and (b) select the most appropriate computational tools to address a particular linguistic problem. This is especially important in forensic contexts, where linguists, in addition to reporting the results of the analysis, need to justify their conclusions scientifically and ensure transparency for court purposes. The boundaries of the concept of computational linguistics are thus blurred, rather than clearly-defined.

The future looks challenging on the computational forensic linguistics front. The development of machine learning techniques, and eventually of artificial intelligence (AI), will raise new issues for forensic linguists. On the computational side, exciting and highly relevant events have been organised. In addition to the PAN competitions over the years, Poleval 2019[3] organised a task aimed at (1) detecting harmful tweets in general, and (2) detecting the type of harm (cyberbullying or hate-speech). The results of the competition will be interesting to see, especially in comparison with the type of analysis usually conducted by forensic linguists. If, on the one hand, AI in particular will be increasingly more competent in producing human-like' texts, on the other (computational) forensic linguists will face the need to develop, test and perfect their methods and techniques to address ever more forensic problems originated by the growing complexity of computer systems. Even if the 'master algorithm' (Domingos, 2015) (one that is able to control all algorithms) is ever discovered, its usefulness in forensic contexts would be very limited: since AI systems operate as black boxes, the results of their analyses cannot be explained – and certainly not to the extent and with the level of transparency required by the courts; yet, they can play a core role in investigative contexts.

Conversely, forensic linguistic expertise will certainly remain crucial, both in cases identical to the ones applicable nowadays, and possibly in other ways of which we are still unaware. If machines are able to generate human-like text, for instance, forensic linguists may need to be able to make a distinction between the texts that were produced by humans and those that were produced by machines. Moreover, forensic linguists may need to assist in cases of machine-generated text, in order to establish whether that text shows some resemblance to the textual production of someone who has control over the system, or on the contrary whether text is machine-generated in order to resemble someone else's text. Similarly, plagiarism analysis and detection will require further research. If machines have the power to generate natural language text, the most serious concern will be not whether the text was lifted from someone else, in whole or in part, or even whether purchased from an 'essay bank', but rather whether it has been produced by a machine.

These are just some of the challenges ahead; there will certainly be many more. Whatever the future holds, however, (computational) forensic linguistics will play a role in it.

Acknowledgements

This work was partilally supported by Grant SFRH/BD/47890/2008 and Grant SFRH/BPD/100425/2014 FCT – Fundação para a Ciência e Tecnologia, co-financed by POPH/FSE.

Notes

[1]http://www.philocomp.net/humanities/signature.htm

[2]https://pan.webis.de/semeval19/semeval19-web/

[3]http://poleval.pl/tasks/task6

References

Abdi, A., Shamsuddin, S. M., Idris, N., Alguliyev, R. M. and Aliguliyev, R. M. (2017). A linguistic treatment for automatic external plagiarism detection. *Knowledge-Based Systems*, 135, 135–146.

Allcott, H. and Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*, 31(2), 211–236.

Amelin, K., Granichin, O., Kizhaeva, N. and Volkovich, Z. (2018). Patterning of writing style evolution by means of dynamic similarity. *Pattern Recognition*, 77, 45–64.

American Civil Liberties Union, (2016). *Statement of Concern About Predictive Policing by ACLU and 16 Civil Rights Privacy, Racial Justice, and Technology Organizations*. Rapport interne, American Civil Liberties Union.

Barrón-Cedeño, A., Gupta, P. and Rosso, P. (2013). Methods for cross-language plagiarism detection. *Knowledge-Based Systems*, 50, 211–217.

Barrón-Cedeño, A. and Rosso, P. (2009). On automatic plagiarism detection based on n-grams comparison. In M. Boughanem, C. Berrut, Soule-Dupuy and J. M. Chantal, Eds., *Advances in Information Retrieval*. Berlin, Heidelberg: Springer, 696–700.

Bowker, L. and Pearson, J. (2002). *Working with Specialized Language: A practical guide to using corpora*. London and New York: Routledge.

Butakov, S. and Scherbinin, V. (2009). The toolbox for local and global plagiarism detection. *Computers and Education*, 52(4), 781–788.

Butters, R. R. (2012). Forensic Linguistics: Linguistic Analysis of Disputed Meanings: Trademarks. In C. Chapelle, Ed., *The Encyclopedia of Applied Linguistics*. Oxford, UK: Wiley-Blackwell.

Carreras, X. and Màrquez, L. (2005). Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling. *Proceedings of the Ninth Conference on Computational Natural Language Learning. Association for Computational Linguistics, Ann Arbor, Michigan*, June, 152–164.

Chomsky, N. (1972). *Syntactic structures*. The Hague: Mouton.

Clark, A., Fox, C. and Lappin, S. (2010). Introduction. In A. Clark, C. Fox and S. Lappin, Eds., *The Handbook of Computational Linguistics and Natural Language Processing*. West Sussex: Wiley-Blackwell.

Clough, P. (2003). Old and new challenges in automatic plagiarism detection. In *National Plagiarism Advisory Service, 2003*, 391–407.

Coulthard, M. (2004). Author Identification, Idiolect and Linguistic Uniqueness. *Applied Linguistics*, 25(4), 431–447.

Coulthard, M. and Johnson, A. (2007). *An Introduction to Forensic Linguistics: Language in Evidence*. London and New York: Routledge.

Coulthard, M., Johnson, A. and Wright, D. (2017). *An Introduction to Forensic Linguistics: Language in Evidence*. London and New York: Routledge.

Coulthard, M. and Sousa-Silva, R. (2016). Forensic Linguistics. In R. J. Dinis-Oliveira and T. Magalhães, Eds., *What are Forensic Sciences? – Concepts, Scope and Future Perspectives*. Lisbon: Pactor, chapter Forensic L.

Cruz, A. F., Rocha, G., Sousa-Silva, R. and Cardoso, H. L. (2019). Team Fernando-Pessa at SemEval-2019 Task 4: Back to Basics in Hyperpartisan News Detection. In *12th International Workshop on Semantic Evaluation (SemEval 2019)*, Minneapolis: Association for Computational Linguistics.

Domingos, P. (2015). *The Master Algorithm: How The Quest For The Ultimate Learning Machine Will Remake Our World*. Harmondsworth: Penguin Books.

Finegan, E. (2010). Corpus linguistic approaches to 'legal language': adverbial expression of attitude and emphasis in Supreme Court opinions. In M. Coulthard and A. Johnson, Eds., *The Routledge Handbook of Forensic Linguistics*. London and New York: Routledge, 65–77.

Franco-Salvador, M., Gupta, P., Rosso, P. and Banchs, R. E. (2016). Cross-language plagiarism detection over continuous-space- and knowledge graph-based representations of language. *Knowledge-Based Systems*, 111, 87–99.

Gales, T. (2015). Threatening Stances : a corpus analysis of realized vs. non-realized threats. *Language and Law / Linguagem e Direito*, 2(2).

Grant, T. (2010). Txt 4n6: Idiolect free authorship analysis. In M. Coulthard and A. Johnson, Eds., *Routledge Handbook of Forensic Linguistics*. Routledge.

Grant, T. (2013). Txt 4N6: Method, Consistency, and Distinctiveness in the Analysis of Sms Text Messages. *Journal of Law & Policy*, 21(2), 467–494.

Grant, T. and MacLeod, N. (2018). Resources and constraints in linguistic identity performance – a theory of authorship. *Language and Law/ Linguagem e Direito*, 5(1), 80–96.

Grieve, J., Clarke, I., Chiang, E., Gideon, H., Heini, A., Nini, A. and Waibel, E. (2018). Attributing the Bixby Letter using n-gram tracing. *Digital Scholarship in the Humanities*, fqy042, 1–20.

Grman, J. and Ravas, R. (2011). Improved Implementation for Finding Text Similarities in Large Collections of Data: Notebook for PAN at CLEF 2011. In *Notebook Papers ofCLEF 2011 LABs and Workshops*, volume 1177, Amsterdam.

Grozea, C., Gehl, C. and Popescu, M. (2009). ENCOPLOT: Pairwise sequence matching in linear time applied to plagiarism detection. *CEUR Workshop Proceedings*, 502(January), 10–18.

Grozea, C. and Popescu, M. N. (2011). The encoplot similarity measure for automatic detection of plagiarism: Notebook for PAN at CLEF. In *Notebook Papers of CLEF 2011 LABs and Workshops*, Amsterdam.

HaCohen-Kerner, Y. and Tayeb, A. (2017). Rapid detection of similar peer-reviewed scientific papers via constant number of randomized fingerprints. *Information Processing and Management*, 53(1), 70–86.

Herring, S. C. (2004). Computer-Mediated Discourse Analysis: An Approach to Researching Online Behavior. In S. A. Barab, R. Kling and J. H. Gray, Eds., *Designing for Virtual Communities in the Service of Learning*. Cambridge: Cambridge University Press, 338–376.

Hoad, T. C. and Zobel, J. (2003). Methods for Identifying Versioned and Plagiarised Documents. *Journal of the American Society for Information Science and Technology*, 54(3), 203–215.

Howard, R. M. (1995). Plagiarisms, Authorships, and the Academic Death Penalty. *College English*, 57(7), 788–806.

Johnson, A. (1997). Textual kidnapping - a case of plagarism among three student texts? *The International Journal of Speech, Language and the Law*, 4(2), 210–225.

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

Kasprzak, J. and Brandejs, M. (2010). Improving the Reliability of the Plagiarism Detection System – Lab Report for {PAN} at {CLEF} 2010. In M. Braschler, D. Harman and E. Pianta, Eds., *CLEF (Notebook Papers/LABs/Workshops)*.

Kay, M. (2003). Introduction. In R. Mitkov, Ed., *The Oxford Handbook of Computational Linguistics*. Oxford: Oxford University Press, xvii – xx.

Kredens, K. (2016). Conflict or convergence?: Interpreters' and police officers' perceptions of the role of the public service interpreter. *Language & Law / Linguagem e Direito*, 3(2), 65–77.

Lum, K. and Isaac, W. (2016). To predict and serve? *Significance*, 13(5), 14–19.

Maia, B. (1997). Do-it-yourself corpora ... with a little bit of help from your friends! In B. Lewandowska-Tomaszczyk and P. J. Melia, Eds., *PALC '97 Practical Applications in Language Corpora*. Lodz: Lodz University Press, 403–410.

Maurer, H., Kappe, F. and Zaka, B. (2006). Plagiarism - A Survey. *Journal of Universal Computer Science*, 12(8), 1050–1084.

McEnery, T. (2003). Copus Linguistics. In R. Mitkov, Ed., *The Oxford Handbook of Computational Linguistics*. Oxford: Oxford University Press, 448–463.

Meyer Zu Eissen, S. and Stein, B. (2006). Intrinsic Plagiarism Detection. In *Proceedings of the European Conference on Information Retrieval (ECIR-06)*, 1–4.

Mitkov, R. (2003). Preface. In R. Mitkov, Ed., *The Oxford Handbook of Computational Linguistics*. Oxford: Oxford University Press, ix – x.

Mozgovoy, M. (2008). *Enhancing Computer-Aided Plagiarism Detection*. Saarbrücken: VDM Verlag Dr. Müller.

Neme, A., Pulido, J. R., Muñoz, A., Hernández, S. and Dey, T. (2015). Stylistics analysis and authorship attribution algorithms based on self-organizing maps. *Neurocomputing*, 147(1), 147–159.

Ng, E. (2016). Do they understand?: English trials heard by chinese jurors in the Hong Kong Courtroom. *Language & Law / Linguagem e Direito*, 3(2), 172–191.

Nini, A. (2018). An authorship analysis of the Jack the Ripper letters. *Digital Scholarship in the Humanities*, 33(3), 621–636.

Oberreuter, G., L'Huillier, G., Ríos, S. A. and Velásquez, J. D. (2011). Approaches for Intrinsic and External Plagiarism Detection: Notebook for PAN at CLEF 2011. In *Notebook Papers ofCLEF 2011 LABs and Workshops*, Amsterdam.

Pataki, M. (2012). A new approach for searching translated plagiarism. In *Proceedings of the 5th International Plagiarism Conference*, Newcastle upon Tyne.

Paul, P. P., Sultana, M., Matei, S. A. and Gavrilova, M. (2018). Authorship disambiguation in a collaborative editing environment. *Computers and Security*, 77, 675–693.

Potthast, M., Eiselt, A., Barrón-Cedeño, A., Stein, B. and Rosso, P. (2011). Overview of the 3rd International Competition on Plagiarism Detection. In V. Petras, P. Forner and P. D. Clough, Eds., *Notebook Papers of CLEF 2011 LABs and Workshops*.

Potthast, M., Stein, B., Eiselt, A., Barrón-Cedeño, A. and Rosso, P. (2009). Overview of the 1st International Competition on Plagiarism Detection. In B. Stein, P. Rosso, E. Stamatatos, M. Koppel and E. Agirre, Eds., *SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09)*, 1–9, Valencia.

Quijano-Sánchez, L., Liberatore, F., Camacho-Collados, J. and Camacho-Collados, M. (2018). Applying automatic text-based detection of deceptive language to police reports: Extracting behavioral patterns from a multi-step classification model to understand how we lie to the police. *Knowledge-Based Systems*, 149, 155–168.

Sarwar, R., Li, Q., Rakthanmanon, T. and Nutanong, S. (2018). A scalable framework for cross-lingual authorship identification. *Information Sciences*, 465, 323–339.

Saunders, J., Hunt, P. and Hollywood, J. S. (2016). Predictions put into practice: a quasi-experimental evaluation of Chicago's predictive policing pilot. *Journal of Experimental Criminology*, 12(3), 347–371.

Shuy, R. W. (2008). *Fighting over Words.* Oxford: Oxford University Press.

Sorokina, D., Gehrke, J., Warner, S. and Ginsparg, P. (2006). Plagiarism detection in arXiv. *Proceedings - IEEE International Conference on Data Mining, ICDM*, July, 1070–1075.

Sousa-Silva, R. (2013). *Detecting plagiarism in the forensic linguistics turn.* , Aston University.

Sousa-Silva, R. (2014). Detecting translingual plagiarism and the backlash against translation plagiarists. *Language and Law / Linguagem e Direito*, 1(1), 70–94.

Sousa-Silva, R., Laboreiro, G., Sarmento, L., Grant, T., Oliveira, E. and Maia, B. (2011). 'twazn me!!! ;(' Automatic Authorship Analysis of Micro-Blogging Messages. In R. Muñoz, A. Montoyo and E. Métais, Eds., *Lecture Notes in Computer Science 6716 Springer 2011*, volume Natural La, 161–168, Berlin and Heidelberg: Springer – Verlag.

Stamatatos, E. (2006). Ensemble-based Author Identification Using Character N-grams. In *Proceedings of the 3rd International Workshop on Textbased Information Retrieval*, 41–46: Citeseer.

Stamatatos, E. (2009a). A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for Information Science and Technology*, 60(3), 538–556.

Stamatatos, E. (2009b). Intrinsic Plagiarism Detection Using Character. In B. Stein, P. Rosso, E. Stamatatos, M. Koppel and E. Agirre, Eds., *SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09)*, 38–46: Universidad Politécnica de Valência.

Stein, B., zu Eissen, S. M. and Potthast, M. (2007). Strategies for retrieving plagiarized documents. In *SIGIR'07*, 825–826, New York, New York, USA: ACM Press.

Svartvik, J. (1968). *The Evans statements: a case for forensic linguistics.* Goteborg: University of Goteborg.

Turell, M. T. (2004). Textual kidnapping revisited: the case of plagarism in literary translation. *The International Journal of Speech, Language and the Law*, 11(1), 1–26.

Turell, M. T. (2010). The use of textual, grammatical and sociolinguistic evidence in forensic text comparison. *The International Journal of Speech, Language and the Law*, 17(2), 211–250.

Vani, K. and Gupta, D. (2017). Text plagiarism classification using syntax based linguistic features. *Expert Systems with Applications*, 88, 448–464.

Vani, K. and Gupta, D. (2018). Unmasking text plagiarism using syntactic-semantic based natural language processing techniques: Comparisons, analysis and challenges. *Information Processing and Management*, 54(3), 408–432.

Weaver, W. (1955). Translation. In W. N. Locke, and A. D. Boothe, Eds., *Machine Translation of Languages*. Massachussets: MIT Press, 15–23. Reprinted from memorandum by Weaver in 1949.

Woolls, D. (2003). Better tools for the trade and how to use them. *Forensic Linguistics*, 10(1), 102–112.

Woolls, D. (2010). Computational Forensic Linguistics: Searching for similarity in large specialised corpora. In M. Coulthard and A. Johnson, Eds., *The Routledge Handbook of Forensic Linguistics*. Milton Park, Abingdon, Oxon; New York, NY: Routledge, 576–590.

Woolls, D. and Coulthard, M. (1998). Tools for the Trade. *International Journal of Speech, Language and the Law*, 5(1), 33–57.

Zhang, C., Wu, X., Niu, Z. and Ding, W. (2014). Authorship identification from unstructured texts. *Knowledge-Based Systems*, 66, 99–111.