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.