Chapter 19 Forensic Linguistics

1. INTRODUCTION

One area of applied linguistics in which there has been increasing trends in both (i) utilising technology to assist in the analysis of text and (ii) scrutinising digital data through the lens of traditional linguistic and discursive analytical methods, is that of forensic linguistics. Broadly defined, forensic linguistics is an application of linguistic theory and method to any point at which there is an interface between language and the law. The field is popularly viewed as comprising three main elements: (i) the (written) language of the law, (ii) the language of (spoken) legal processes, and (iii) language analysis as evidence or as an investigative tool. The intersection between digital approaches to language analysis and forensic linguistics discussed in this chapter resides in element (iii), the use of linguistic analysis as evidence or to assist in investigations. Forensic linguists might take instructions from police officers to provide assistance with criminal investigations, or from solicitors for either side preparing a prosecution or defence case in advance of a criminal trial. Alternatively, they may undertake work for parties involved in civil legal disputes. Forensic linguists often appear in court to provide their expert testimony as evidence for the side by which they were retained, though it should be kept in mind that standards here are required to be much higher than they are within investigatory enterprises.

Forensic linguists find themselves confronted with a wide range of tasks, including settling matters of disputed meaning of slang terms (Grant, 2017), commenting on the level of influence one speaker may have had over the contributions of another (Coulthard, 1994), determining the linguistic proficiency of an individual (Cambier-Langeveld, 2016), or providing an opinion on whether or not an interviewee understood the content of what her/his interviewer was saying (Brown-Blake and Chambers, 2007) One of the more common linguistic problems that forensic linguists are faced with is identifying the author(s) responsible for writing a text or texts, a process widely referred to as 'authorship analysis'. Over time, as the value that forensic linguists can bring to criminal (and civil) investigations has become clear, those in law enforcement have enlisted their assistance in the gathering of evidence and, in some cases, have provided linguist-led courses to trainee officers. For the most part, these activities are underpinned by scholarly attempts to refine the methods informing such work, and standards and reliability in forensic work have found themselves at the centre of many a scholarly debate in the field (see Butters, 2012).

This chapter explores the interface between forensic linguistics and digital methods in the collection and analysis of linguistic evidence, particularly in relation to questions of authorship. We describe the different processes through which digital approaches have become embedded in the forensic investigation of language use, and by drawing on two ongoing projects, we cast light on the empirical potential afforded by digital methods, tools and texts to researchers seeking to improve the standard of evidence and investigative assistance that forensic linguists can offer.

2. BACKGROUND

Research in the field of authorship analysis has tended to concern itself with developing methods for correctly attributing authorship of lengthy literary texts to one of a small closed set of candidate authors (see Koppel *et al*., 2011). Addressing the question of authorship tends to involve the comparison of the 'anonymous' text with known sample writings of the candidate authors, using statistical or stylistic methods, or a combination thereof. The reality of forensic casework, however, presents several challenges not addressed by the vast majority of these scholarly investigations into authorship analysis questions. Firstly, the volume of known writing of a suspect that is available to the linguist may be rather limited; it may be of a different genre to the anonymous text; and it may have been produced by different means (a desktop word processor versus a mobile phone, for example). A second set of problems are those that arise from the lack of any knowledge of base rates for linguistic features considered to be individuating, or part of an author's *idiolectal style* (Turell and Gavaldà, 2013). That is, there is often insufficient appropriate data available to provide us with the general frequency of the linguistic features or patterns in any given population. Therefore,while we may be able to describe a particular linguistic choice as 'marked', how distinctive it actually is in a given population is arguably unknowable. We discuss proposed methods for addressing these shortcomings later in the chapter.

For the forensic authorship analyst, the detailed and systematic description and analysis of an individual's linguistic choices when constructing texts is a point of departure.. In this chapter, we are concerned with two investigative tasks that this might feasibly feed into: the attribution of anonymous texts to the correct author, and the synthesis of a particular persona for investigative purposes. There has been a steady rise in the proportion of forensic linguistic cases involving digital communications (Coulthard *et al*. 2011: 538), ranging from cases involving disputed authorship of SMS text messages (Grant 2013) or e-mails (Turell 2010; Coulthard 2013) to ascertaining meaning in Instant Messaging (IM) conversations (Grant, in press). In line with this development, more scholarly attention has also been paid to refining methods for authorship analysis of electronic messages to underpin such case work (see, for example, Abbasi and Chen, 2008; Koppel *et al.*2002; Silva *et al.,* 2011). Methodological frameworks have evolved in line with requirements for them, as the work of the forensic authorship analyst has moved increasingly further into the digital realm.

3. CRITICAL ISSUES AND TOPICS

International jurisdictions impose various scientific 'standards' which expert evidence must satisfy in order to be admissible in court, such as the Criminal Practice Directions Amendment No.2 [2014] in England and Wales and the Daubert Criteria in the United States. In light of these standards, the field of forensic linguistics is currently experiencing a period of intense self-reflection, and debate abounds about how best to ensure that practitioners work to the necessary scientificstandards., The International Association of Forensic Linguists (IAFL) published its Code of Practice in 2013, reflecting a concern among its membership that the organisation should formally commit to upholding minimal standards in casework. It is important to keep in mind that this merely serves to provide guidelines to linguists wishing to undertake casework and the IAFL has no enforcement capacity in relation to the Code. In this chapter we identify two challenges currently facing forensic linguistics, in which digital approaches and methods can be instrumental in ensuring that standards of reliability are met and the delivery of justice improved: (i) the rise in cyber-enabled crime, and (ii) improving the reliability of analysis and evidence in forensic authorship cases., First, we outline the current state-of-play with regard to these challenges in the field. Then, we describe two separate programmes of work being undertaken to address these challenges. Both of these case studies demonstrate how methods within digital humanities are facilitating linguistic analysis in forensic contexts that would simply not be possible without methods of data collection and analysis afforded by digital approaches. We conclude the chapter by reflecting on the digital humanist nature of the methods described, and comment specifically on the computer-human relationship central to them both.

3.1 The need for digital methods: the rise in cyber-enabled crime

Sociolinguistics has evolved at much the same rate as other areas of the humanities in terms of the upsurge since the late 1990s in research addressing or resting upon digital data and/or methods (see, for example, Chapter 14 of this volume; Seargeant and Tagg, 2014). Reflecting a trend across linguistics in general, recent scholarly interest in computer-mediated discourse has been characterised by the use of mixed methodologies, incorporating digital methods of analysis alongside those perhaps more traditionally associated with their discipline (see Bolander and Locher, 2014). The fact that individuals engage in socially meaningful interactions online in a way that typically leaves a textual trace and is thus accessible to the researcher’s scrutiny (Herring, 2004) has made such communications a site where empirical methods have been put to work shedding light on a wide range of interactive phenomena. Contemporaneously, these methods have themselves seen substantial development as software has (semi-) automated and sped up analysis in many sub-fields of the discipline. Larger collections of linguistic and multi-modal data can now be subject to a level of detailed inquiry not possible before, and the overall processes of collecting, storing, accessing, navigating and coding data have been accelerated immeasurably by technological advancements including corpus programs, annotation tools, automated taggers, and qualitative analysis suites. Alongside these advances, however, a more sinister side to technology has also been evolving. Increasingly, forensic linguists are approached with digitally produced texts, and often, such as in the case of texts containing conspiracies or grooming material, the production of those digital texts is, in itself, a crime.

The UK Home Office makes a distinction between ‘pure’ cybercrime, i.e. attacks on digital systems, and cyber-*enabled* crime, i.e. offences traditionally committed in offline contexts which can increase in their scale or reach by being perpetrated via the internet (McGuire and Dowling, 2013). There is no doubt that child abuse is one area of criminal activity that has been made easier and less risky by technological advances. The sexual grooming of children, i.e. the preparation of children for sexual activity (Carmody and Grant, in press), is a widespread issue, and one that has escalated in line with the advancement of the world wide web. Increased access to large numbers of like-minded individuals *and* potential victims at the click of a button has led to figures suggesting that 60% of children in the UK have been sexually solicited online (Internet Watch Foundation, 2013). Compounding these statistics, the anonymity afforded by the internet means decreased levels of perceived? risk involved in such activities. The rise of the dark web, a heavily encrypted and thus anonymous means of accessing online content, has rendered traditional methods of offender identification, such as tracking geolocation and IP addresses, wholly ineffectual. Thus, online policing of child abuse has been described as being in crisis, with undercover operatives in dire need of alternative methods for pursuing the identification and prosecution of offenders.

This is a task with which the forensic linguist is well equipped to assist. The situation highlights the need for effective, well informed, evidence-based approaches to policing the digital sphere, and methods that have been developed for authorship analysis can easily be adapted to this problem. Technological advances afford opportunities to the linguistic researcher, and thus the investigator, just as they do the online criminal. Just as style markers are used to attribute authorship of unknown texts (see section 5.1), they are also put to good use in describing the linguistic persona of an individual (such as a victim) to the extent that that persona can be assumed by another individual (such as an Undercover Officer (UCO)). There is thus an obvious demand for the talents of a forensic linguist among law enforcement professionals seeking to successfully assume identities online.

3.2 Reliability of evidence

Today, most research in forensic authorship analysis is informed, either explicitly or implicitly, by the criteria for admissibility of evidence in court. This has predominantly been in the context of the United States and the Daubet criteria (Daubert vs. Merrell Dow Pharmaceuticals, Inc [1993]), which were established to ensure that expert evidence offered is 'scientifically valid', that is:

> 1. Whether the theory offered has been tested;
>
> 2. Whether it has been subjected to peer review and publication;
>
> 3. The known rate of error; and
>
> 4. Whether the theory is generally accepted in the scientific community
>
> (Solan and Tiersma 2004: 451)

There are a number of studies which discuss authorship methods, and the practices of forensic linguists more generally, in relation to these criteria (e.g. Solan and Tiersma 2004; Coulthard 2004; Howald 2008; Juola 2015). Similar criteria are set for expert evidence in England and Wales, the 'reliability' of which is assessed on the basis of factors including: the quality and completeness of the data used, the accuracy or precision rates of the method employed, the extent to which the material or method on which the expert relies has been peer-reviewed, and whether the expert's method followed established practice in the field (Crown Prosecution Service, 2014). Such criteria now underpin much of the research undertaken by authorship analysts, as they reflect on how the methods they are developing would fare when tested against these parameters.

At the same time, the regulation of covert policing techniques is plagued by numerous ethical issues, making it fertile ground for scholarly endeavours by security ethicists (e.g. Nathan, in press). Central to these concerns is the potential for accusations to be levelled at UCOs of acting as *agents provocateur* (AP), or of entrapment – circumstances where a person has been induced to commit an offence which he or she would not have committed but for the inducement. The practice of assuming alternative identities in order to draw out offenders and secure an arrest is fraught with ethical difficulties in this regard, as reflected in judicial vigilance around the issue – compliance with Home Office guidelines is pivotal to the success of a prosecution. Modern approaches to establishing whether or not entrapment has taken place require consideration of whether there was reasonable grounds to suspect an offence was to be committed, as well as the police activity being properly authorised and supervised (Ellison and Morgan, 2015). Thus, accurate record keeping forms an important part of preparing and carrying out an identity assumption task, and a UCO's digital pocketbook is her/his best defence against any potential allegations of impropriety. There is a place here for the forensic linguist – as we shall move on to discuss, one aspect of a victim's linguistic style that a UCO may wish to emulate is that of topic management, in a situation where any mention of sexual topics might be viewed as an enticement. If, however, a UCO is able to demonstrate that the initiation of sexual topics was a stylistic marker of the victim's linguistic persona, then such accusations may be mitigated. A thorough analysis of all facets of an individual's linguistic identity is therefore crucial in ensuring best practice and the highest chance of successful prosecutions.

4. MAIN RESEARCH METHODS

4.1 Stylometry

Stylometry is the umbrella term used to refer to the wide range of methodological approaches to authorship analysis in which the similarity or difference between authors' styles is statistically measured on the basis of their use of a particular set of linguistic features (Coulthard et al., 2017: 153). Stylometry is used to address three different authorship problems: authorship attribution, which involves identifying which author from a set of candidates is most likely to have written an 'anonymous' text, authorship verification, which involves determining whether a particular author is responsible for a particular text, and author profiling, which is the prediction of an author's social characteristics (age, gender, personality type, occupation, ethnicity etc.) on the basis of language use. Stylometry has developed primarily from a mathematical tradition and the most seminal early study in stylometric authorship analysis is the work of two statisticians: Fredrick Mosteller and David Wallace (1964). In an attempt to attribute a set of twelve disputed *Federalist Papers* to one of two candidate authors—Alexander Hamilton and James Madison-- Mosteller and Wallace compared the disputed texts with known writings of the two authors on the basis of how frequently they used 70 high-frequency common grammatical words, including determiners, prepositions and conjunctions. They found significant differences in usage between the authors, and they assigned all twelve disputed texts to Madison, a conclusion which supports the prevailing opinion of historians. *The Federalist Papers* study was somewhat of a watershed moment for stylometric authorship analysis, and since then stylometry has benefited greatly from the increased methodological opportunities afforded by digital and technological advances.

Increased digital capabilities have allowed for the collection of far larger datasets than before. Corpora of newspaper articles, blog posts, emails, Instant Messages, literary texts, SMS messages, social media posts, and scientific texts, often running into the millions of words, are now regularly the objects of stylometric techniques. As for the methods themselves, they too have grown alongside increases in computer processing powers. Koppel et al. (2008) and Stamatatos (2009) provide comprehensive surveys of the linguistic features and statistical procedures that are commonly used in stylometric authorship work. The linguistic features—or ‘style markers’— on which similarities between texts and authors are based, range from vocabulary richness measures and function/content word frequencies to character n-grams and word sequences. Beyond lexical features, syntactic variation between authors has been captured by part-of-speech n-gram frequencies and some studies have compared authors on the basis of semantic features such as synonym preferences and types of speech acts used. When it comes to the statistical means by which similarity between texts and authors is measured, Koppel *et al.* (2013: 318-9) distinguish between those in the ‘similarity-based’ paradigm, and those in the ‘machine learning’ paradigm, for addressing the simplest kind of authorship problem in which the true author of a document is assumed to be one of a small number of candidates. Similarity-based approaches involve the use of a given metric to measure how similar an anonymous document is to the known texts in terms of the linguistic features under investigation, and the anonymous document is attributed to that author whose known writing is most similar. In the machine learning paradigm, the set of known writings of each candidate author is used to train a classification algorithm that can assign the documents to their correct authors. This classifier is then used to estimate the author of the anonymous document.

Stylometric approaches as outlined here are applications are a cornerstone of digital humanities (e.g. Jockers and Underwood, 2016), but they have primarily been developed in non-forensic contexts. However, recent years have seen an increased focus on the forensic applications of such techniques to combat cybercrime, and how they can assist in the exploration of areas including extremist web forums (Abbasi and Chen, 2005) and drug trafficking forums (Rico-Sulayes, 2011). It is often held that stylometric approaches, which are tested in experimental conditions, have known error rates and involve less human interference, provide more reliable and objective results and so are best equipped to satisfy the criteria for the admissibility of evidence. However, there are some counter-arguments that stylometric approaches and their results are too difficult to present and explain to lay jurors and judges, either because there is no theoretical motivation for choosing particular style markers for analysis (Argamon and Koppel, 2013: 301), because they generally require large datasets to be effective (Koppel et al. 2011) or because they are based on complex statistical assumptions with which it is unrealistic to expect the jury to engage, leading ultimately to the evidence being distorted rather than illuminated (Cheng, 2013: 547). Therefore, although the digital tools are effective and well-tested, we mustn’t lose sight of the role of the human analyst.

4.2 Stylistic approaches

Stylistic approaches contrast with stylometric approaches in that rather than relying on computers or algorithms, it is the analyst themselves who compares the documents. Again, in the most straightforward authorship problem, this involves the expert performing a systematic linguistic analysis of the known writings available for the candidate author(s), and identifying a number of markers which they use with some level of *consistency* and which appear to be characteristic of their style(s). If a case involves more than one possible author, then they must have observably *distinctive* styles for the analysis to continue. Attention then shifts to the anonymous document(s) to identify whether the styles exhibited within them are consistent with the known writings of any of the candidate authors. Such stylistic methods are better suited, or indeed are the only option available, when the amount of known and/or anonymous data is scant. There are also arguments that stylistic analyses are more firmly grounded in theories of linguistic variation than their stylometric counterparts (Nini and Grant, 2013: 176), and that qualitative evidence is more easily demonstrable than statistical results to juries, who are more comfortable weighing up this kind of evidence (McMenamin, 2002: 129; Cheng 2013: 547). However, stylistic approaches have been heavily criticised in recent years, primarily on the basis that they are too heavily reliant on the analyst's subjective judgements. Chaski (2005: 2) for instance, states that 'without the databases to ground the significance of stylistic features, the examiner's intuition about the significance of a stylistic feature can lead to methodological subjectivity and bias'. As a result, when such subjectivity is used as the foundations for stylistic contrasts drawn between texts and authors, 'these are sometimes loosely defined and can be harder to measure and evaluate' (Nini and Grant 2013: 176).

Stylometric and stylistic approaches are at best divergent, and at worst competing, with nobody knowing for sure which would fare best in any given case (Solan 2013: 557). A small number of recent studies have made promising moves in explicitly combining the stylistic and statistical (e.g. Grant 2013; Nini and Grant 2013), but one thing remains clear; stylometry has flourished as a digital humanity (see Juola 2015), while stylistic methods are yet to embrace the opportunities afforded to them by technological advancements. The area which can most obviously address this is the increased use of digitally-collected corpora to serve as, in Chaski's (2005: 2) terms, databases in which to 'ground the significance of stylistic features', and this is demonstrated in Section 5.1.

4.3 Analysing Computer Mediated Discourse

Other forensic linguistic work is concerned less with attributing or verifying authorship of texts and more with providing a descriptive account of an individual's linguistic identity such that it can be successfully assumed by another individual (see Section 5.2). Such work is often designed from the outset to have direct impact on investigative training and practice. As such, the data are often tackled in such a way as to yield results that could easily be incorporated into current training, as well as in this case providing a point of departure for the development of software to semi-automate the process of analysing a linguistic persona in preparation for identity assumption. As mentioned above, reliability of forensic linguistic assistance offered to investigators is not required to be of the same standard as that submitted as evidence. Findings from the project outlined below are intended only to relieve some of the investigators' burden, rather than to provide evidence at a criminal trial.

In the work described in Section 5.2 below, observations about experimental participants' language use through the medium of Instant Messaging (IM) provide the grounding for interpretations of the relationship between language use and online identities. The focus on IM as the principle means through which these identities are discursively projected and maintained justifies the selection of Computer Mediated Discourse Analysis (CMDA) (Herring, 2004; 2012) as a starting point for the analyses. Rather than referring to an approach in itself, CMDA refers rather to an organisational principle - the fairly straightforward transferal of methods from linguistics and 'traditional' discourse analysis across to digital media. CMDA allows questions of broad social significance to be subject to robust fine grained empirical investigation through a range of analytical methods, which simultaneously feeds in to a rigorous evidence based approach to a particular set of investigative problems.

Herring (2004) advocates the use of a 'tool-kit' approach to CMDA, within which the researcher selects analytical tools relevant to the questions they are seeking to address. The four-level hierarchy of CMDA begins at the micro-linguistic level of structure and moves up through analyses at the level of meaning, such as scrutinising the data through the lens of speech act theory (Searle, 1969; 1975), to interaction management such as the analysis of turn taking patterns and topic control, and finally to the macro-level of social phenomena. Most studies of computer-mediated discourse have tended to take phenomena at just one of these levels as their focus. In order to be maximally forensically relevant, however, all four were incorporated to some degree into the analyses that formed part of the project described under 5.2 below. At the *structural* level choices relating to lexis, punctuation and graphology are attended to, while at the level of *meaning* turns are categorised according to the speech act(s) they perform. At the level of *interaction*, topics that are introduced, maintained or rejected by an individual are recorded, along with their preferences as far as turn length, turn structure, openings and closings are concerned. Identities, ostensibly a feature at the *social* level, are best seen as products of the resources available to them at all three of the prior linguistic levels. The linguistic nature of identity is a particularly salient consideration in online contexts, since many of the resources on which we ordinarily rely for the production and interpretation of identities tend to be absent (Donath, 1999). Style markers at all levels can be used to characterise a linguistic persona, and this composite style can then theoretically be selected for performance by another individual.

As we discussed in Section 3.1 above, technological advances now allow for the speedy and thorough analysis of large amounts of linguistic data on an unprecedented scale. The data in this study were coded with the assistance of QSR NVivo, a piece of software designed to help organise and find rich insights into unstructured qualitative data. The application of these digital methods to often 'messy' data ensures that the data are easier to navigate and far less challenging to manage. Overviews can be generated of each individual's preferences at the three linguistic levels described above across the duration of their IM conversations. Comparisons of these overviews then elucidate the most marked differences between participants, facilitating an examination of how successful individuals are at performing particular style markers that are not ordinarily part of their own linguistic repertoire. We expand on this process in Section 5.2.

5. CURRENT CONTRIBUTIONS AND RESEARCH

We report here on two substantial research projects, both of which highlight the benefits of applying analytical methods originating from discourse analysis and sociolinguistics to substantial collections of digital texts in order to answer forensically relevant questions. The studies also demonstrate some ways in which developments in digital methods have allowed for unprecedented exploration of large volumes of data. The works are presented here in order to showcase the breadth of research carried out under the auspices of language as evidence, both as a subject of scholarly attention in general, but also as a means of reinforcing methods for addressing such questions in the 'real world'. One project does this with a focus on authorship analysis and the potential for linguistic markers to distinguish individuals from one another to the extent that they can be used to assign texts of unknown authorship to the correct author. The other does so with a view to authorship synthesis, the process by which individuals incorporate features of a target persona's language use into their own for the purposes of identity assumption.

5.1 Using corpora as relevant population data: an Enron case study

While most forensic linguists have generally dismissed comparisons between their evidence and that of DNA (or fingerprinting) (e.g. Coulthard, 2004: 432), primarily because of the variability of linguistic evidence, stylistic analysis stands to benefit from following some of the DNA profiling process. In particular, the use of population data to measure the frequency or rarity of particular genetic patterns can be borrowed by forensic linguists. Large scale corpora can be used as population data to test the frequency or rarity of a particular linguistic feature or variant identified in an analysis. Such a process has the potential to address the major criticisms levelled at stylistic approaches to authorship analysis: rather than the diagnostic power of a particular style marker being determined by the analyst's intuition, population data could be used to measure the evidential value of finding that style marker in two separate writings.

This is not a new idea. Coulthard (1994) used a corpus approach to identify the marked use of the word *then* in the famous case of Derek Bentley's disputed statement to the police. However, the corpora that he used were small by modern standards: 2,270 words of other witness statements and 1.5 million words of the Corpus of Spoken English (a subset of the COBUILD Bank of English). Since then, many others have championed the use of corpora in stylistic authorship analysis. One may expect the mantra of 'bigger is better' to hold in discussions of 'population' data – the more data that is available for comparison, the more reliable the tests of frequency or rarity. Therefore, the availability of general reference corpora such as the British National Corpus (100m words), the Corpus of Contemporary American English (560m words and growing with additions each year) and the Global Web-Based English corpus (1.9b words) may seem tantalising in this regard. However, the appetite amongst forensic linguists is not just for population data, but *relevant* population data. What 'relevant' means is up for debate. For some, a relevant population would comprise language users from the same 'linguistic community' (Turell and Gavaldà, 2013: 499). For others, genre must be consistent with the anonymous texts under examination (Grant, 2013: 473), while for Kredens

(2002: 435), reliable relevant population data should be characterised by biological, social and interactional variables identical with those of the anonymous document(s) of the case. Therefore, large scale reference corpora are not relevant enough. It is not sufficient, for instance, to compare the linguistic features exhibited in an anonymous tweet with millions of words of newspaper articles or fiction. Rather, *specialised* corpora, complied according to consistency of entries with regards to parameters of setting, purpose, genre, discourse type and variety (Flowerdew: 2004: 21) are required 'to provide population-specific statistics on usage' (Kredens and Coulthard, 2012: 507).

To date, this appetite for relevant population data amongst forensic linguists has not been matched by their uptake of it. This is likely due to the time and financial cost of compiling and organising such corpora. This is especially true if the corpora are to be built for research purposes only, rather than benefiting a particular case, as the relative pay-off for the work may seem less. That is, linguists may be less inclined to expend resources on building a corpus if it is not to immediately serve their analysis in a specific case on which they are working. However, Turell (2010: 212) emphasises the role that research studies have in developing methods and assisting the expert witness, and so the constructing and testing of relevant reference corpora, rather than simply wishing for them, is an important endeavour. Coulthard (2013: 466) states that 'forensic linguists are never going to have reliable population statistics to enable them to talk about the frequency or rarity of particular linguistic features'. Fortunately, with the development of digital humanities and related technologies, the collection, storage and analysis of specialised corpora as relevant population data is far easier than it has been in the past, and there is cause for optimism. Examples include the development of Application Programming Interfaces (APIs) and web-crawling techniques which have made it possible to collect millions or billions of online text data such as web forum posts (e.g. Törnberg and Törnberg (2016), blogs (Schler et al., 2006) and tweets (Grieve et al., 2017) and blogs (2017). Such technologies facilitate the collection of specialised corpora of various kinds on a scale unimaginable when Coulthard (1994) wrote about corpora in authorship analysis, and stand to greatly benefit authorship cases involving certain text types, should they be embraced by forensic linguists.

Two studies which use such corpora to demonstrate how such data can be used in an authorship context are Author B (2013) and Author B (2014). Both studies use the Enron Email Corpus to test the frequency or rarity of authors' stylistic choices. In 2003, as part of the Federal Energy Regulatory Commission's (FERC) legal investigation into Enron's accounting practices, the email data of some Enron employees was made publicly available online. Cohen (2009) formatted this data and made it suitable for research, and it is his version of the corpus that Author B (2013) and Author B (2014) made use of. In order to optimise the corpus for authorship analysis purposes, the data had to be cleaned in various ways, including the removal of duplicate emails, removal of email conversation threads, retaining only *sent* emails, and the removal of unwanted metadata carried over from the original FERC set. This data clean-up was performed programmatically by David Woolls of CFL Software Ltd. The Enron Email Corpus, which after cleaning contains 63,369 emails, sent by 176 authors totalling 2,462,151 tokens, can be considered as 'relevant population data' for a number of reasons. . It is 'relevant' in that it pertains to the *linguistic community* of Enron employees, or perhaps it is more appropriate to say that it represents the communication of a specific Community of Practice (Eckert and McConnell-Ginet 1998: 490). Furthermore, both men and women authors are included, all of the texts in the corpus are of the same text type or genre (email), they are all written within a four year period (1998-2002), and there is some control of register with portions of the employees sharing the same occupation-related lexis. This means that when the rarity or frequency of particular style markers is being tested, Enron emails are being compared with Enron emails, and all of those in the population are writing in the same medium, in the same community of practice, working for the same company at the same time and many have the same job.

Author B (2013) focused on authors' distinctive uses of email openings and closings, which are thought to be as habitual as choices in vocabulary and syntax (Corney et al. 2001). In the first instance, the opening and closing preferences of a small subset of four male traders were identified. Some forms were common across some or all authors, such as the omission of a greeting or farewell altogether, or the use of first name only to sign off. Others, however, were found to distinguish between the four authors. For instance, one of the authors, John Arnold, used the greeting *hello:* which none of the other three did:

Extract 1: John Arnold email

```
<Date: Wed, 8 Nov 2000 09:07:00 -0800 (PST)>
<From: john.arnold@enron.com>
<To: kendrick.brown@eia.doe.gov>
<Subject:>

Hello:
I am not able to pull up the link for the short term outlook
for natural gas. Can you please make sure the link is
updates.
Thanks,
John
```

This greeting form, therefore, could be used to distinguish between this small set of authors. Transposed onto a real authorship case, this would be a style marker which characterises Arnold's style in contrast to that of his three colleagues. However, Arnold did not use this form particularly frequently; it only appeared in seven of his 632 emails (1.11%). Therefore, the evidential strength of this feature as being indicative of his style is not particularly convincing. However, when the frequency of this feature was tested in the relevant population data (i.e. the rest of the Enron corpus) it was found to be used only once by one other author in 40,236 emails (0.002%). In other words, if an email in the Enron corpus contained the greeting *hello:*, it is 555 times more likely that this email belongs to Arnold than anyone else in the population (1.11%/0.002%). In Grant's (2013) terms, it is distinctive of Arnold's writing at 'population level'. Although Arnold is not a frequent user of this greeting, comparing his usage against that of a relevant population provides strong evidence of its rarity, and therefore the likelihood than an email containing that form belongs to him.

Using the Enron corpus as relevant population data in the same way, Author B (2014) investigates the distinctiveness of collocation patterns. Launching from the assumption that collocations are unique to individuals (e.g. Hoey 2005: 181), Author B examines the preferences of one individual, James Derrick, in the use of *please* mitigated directives. Occurring with a frequency of over eleven thousand, *please* is the second most key word in the Enron corpus when compared against the Corpus of Contemporary American English (*thanks* is most key, appearing 14,865 times).

This indicates that the Enron dataset is representative of a community of practice in which interlocutors are generally linguistically polite to one another, and also suggests that requests or mitigated directives are this community's principal speech act (Author B 2014: 47). Nevertheless, an analysis of Derrick's *please* collocates finds that his use of *please* is distinctive. First, some of his most frequently occurring *please* initial collocations are not used by any of the other Enron employees in the population. For example, there are 109 instances of *please* in his emails, seven of which are followed by *format and*. By contrast, *please format and* is not used anywhere else in the Enron population; it is completely distinctive of his idiolectal style. Some of his preferred uses, such as *please print the,* are found with some frequency in the emails of other authors. This three-word sequence accounts for 35 of Derrick's 109 instances of *please* (32.11%). Besides Derrick's uses, there are 10,952 instances of *please* in the remaining Enron population, and 32 (0.29%) of these are part of the longer sequence *please print the*. This means that Derrick is 111 times more likely to use *please print the* than anyone else in the population (32.11%/0.29%). These findings show that even within very common communicative behaviours—*please* mitigated directives—we are able to identify distinctive patterns. Author B (2017) develops this idea further, using the 'relevant population data' approach to demonstrate that different authors in the Enron corpus, with the same job, when faced with the same communicative situation, encode and express the same speech act in different ways, and that the resultant linguistic output can be distinctive enough in the population to correctly identify or verify the author of a questions text. As some readers will have probably already noted, *please format and* and *please print the,* as well the *hello:* greeting in Author B (2013) are rather unremarkable linguistic features. They are far from the ideal 'smoking guns' of authorship like a bizarre spelling of a word or a marked syntactical arrangement. Instead, these features might easily sail under the radar of forensic analysts stylistically examining an anonymous text (or a known writing). Having access to relevant population data allows us not only to identify them as potential markers of authorship, but to reveal that they are in fact *very distinctive* markers of authorship within that population, and that their appearance in any two emails is a strong indication that these emails are written by the same person.

5.2 Assuming identities online

The second study to be described in depth here set out to bring together the numerous facets of individual identity and explore the notion of idiolect in computer mediated discourse. It did so from the theoretical position that identities are actively constructed for particular occasions; that they comprise a fluctuating body of resources from which individuals draw in order to conduct their social business; and that, in the online context at least, they are entirely linguistically and discursively constructed.

The research was designed with the needs of undercover police officers (UCOs) in mind, with a particular focus on investigations into online child sexual abuse as set out in Section 3.1. In this context, operatives are frequently required to assume an alternative identity online, and to sustain this identity in interactions with the perpetrator in order to set up a meeting so an arrest can be effected. The impetus for the work was the researchers' prolonged involvement in the training of UCOs for this task, and a realisation that both the training content and the identity assumption process itself had much to gain from rigorous academic attention and intervention. Thus, a study was carried out with the central aims (i) to assess what linguistic analysis is necessary and sufficient to describe an online persona so that it can be assumed by another individual and (ii) develop theoretically how different individuals establish and perform their online personas through a combination of interactional, topic and stylistic choices.

The focus was on Instant Messaging (IM) as a medium, a type of computer-mediated communication 'involving two parties and done in real time (synchronously)' (Baron, 2013). Communication is facilitated through written exchanges, and, like many other types of Computer Mediated Communication, IM combines qualities typically associated with writing – such as lack of a visual context and paralinguistic cues, physical absence of interlocutors – with properties of spoken language, such as immediacy, informality, reduced planning and editing, and rapid feedback (Georgakopoulou, 2011). IM has thus been described as a 'hybrid' register (Tagliamonte and Denis, 2008). The study draws on the logs of IM conversations between (i) online sexual groomers of children and their victims; (ii) undercover officers (UCOs) posing as individuals involved in sharing indecent images of children and their criminal targets; (iii) police trainers and trainees in simulated online grooming investigations; and (iv) participants in experiments designed with the express purpose of investigating online linguistic identities (see x and Author A, 2016 and Author A and x, 2016 for more on the benefits of using experimental work in research of this nature). Since they are neutral and non-triggering in content, and comprise a more complete set of data than do any of the genuine police logs, it is these experimentally 'got up' IM conversations that form the basis of the discussion here. The tasks were designed with the aim of assessing what level of linguistic analysis is necessary and sufficient for the successful substitution of one interlocutor with another.

Computer mediated communication, like all areas of language use, is multi-faceted, and a number of relevant levels of analysis can be identified as set out by Herring (2004) and described in 4.3 above. Each participant's contributions to (i) IM conversations in which they were not under any instruction to perform an alternative identity and (ii) IM conversations in which they had been tasked with assuming a particular individual's identity, with varying levels of preparation, were analysed at the linguistic levels of *structure, meaning* and *interaction*. Thus, a detailed picture of an individual's *usual* preferences in terms of a wide range of potential style markers from the micro-level, e.g. initialisms, syllable substitution etc. (see Author A and x, 2012 for the preliminary version of the taxonomy that informed this analysis) through to patterns of turn-taking and speech acts, was recorded as well as the extent to which these choices changed when they were tasked with impersonating someone else. The IM logs were not the only set of data to emerge from the experimental work. While at any one time some participants were tasked with taking on alternative linguistic identities, others were tasked with spotting when their interlocutor in an IM conversation was replaced by someone else impersonating them. Information that was noted down by participants as they prepared for the identity assumption task were available for scrutiny, as were participants' musings on why they believed the individual they were talking to was not who they implicitly claimed to be. These documents, alongside the IM logs themselves, formed an important part of the analyses.

The overwhelming majority of participants' comments about what led them to believe their interlocutor had been replaced with someone else related to the structural level of language: features of vocabulary, spelling, spacing, punctuation, and so forth, are what gave people away. The second highest total represented comments relating to the interactional level. Comments in this category relate to changes in speed and pace of typing, length and breaking of turns, and topic management. Closely following behind this level was the social level, which included comments related to identities and relationships. It is crucial to keep in mind that these types of comments are virtually impossible to disentangle from observations at the one or more of the first three linguistic levels – an individual described by a Judge as 'chatty', for example, might receive this description because they produce longer or more turns than other participants (a feature at the interactional level), or an individual may be described as 'bossy' because they issue a high proportion of directives (a feature at the level of meaning). Comments actually at the level of meaning were extremely scarce. It is a fairly straightforward matter to feed this information into training UCOs – a primary focus on low-level linguistic features will stand one in good stead for the identity assumption task, since inconsistencies at this level are the ones most easily spotted by the individual one is attempting to deceive. UCOs' attentions should also be focussed at the interactional and social levels, and, according to these results, minimal attention is required at the level of meaning (although see above). Notes made in preparation for the identity assumption task tell a similar story to these observations. Almost 70% of observations about a target language style were related to potential markers at the structural level.

One important outcome of this project has been Identik, a software tool that partially automates the process of linguistic analysis of an individual's identity, expediting the preparation process for UCOs tasked with identity assumption in live investigations. The tool provides a linguistic summary of each person involved in a pre-loaded IM conversation, providing numerical information about features such as their spelling and punctuation choices and average turn length. When there are two or more variant spellings of an item, the tool provides a ratio of the relative frequencies. The tool also provides a number of free text areas in which UCOs can record their observations about speech acts and openings, among other things. One further feature of Identik is an auto-stylisation function, within which a UCO can input a phrase that is then 'translated' into the language of the selected individual. While not intended to replace the process of human linguistic analysis and identity performance, it is hoped that the tool will relieve some of the cognitive burden on officers engaged in these activities. This project thus represents a tangible impact of linguistic research on policing practice via digital methods.

6. FUTURE DIRECTIONS

There can be no doubt that the digital trend is set to continue unabated, and that forensic linguists will continue to be called upon to provide their assistance with either digital texts, digital investigation methods, or a combination of the two. Both of the case studies that we have discussed here sit firmly within the framework of digital humanities. What is important with both approaches described here is that they combine the efficient use of computer and digital technologies with effective human interpretation and communication of results – a central tenet of 'humanities computing' (Unsworth, 2002). Although the computer is the apparatus through which the forensic analyses described here are performed, the both require careful input, interaction and explanation by the human analyst. For instance, deciding the currency to be attached to discovering that a particular style marker 500 times more likely to be used by one author than another is a task for the analyst. More obviously still, the assumption of identities described here is performed by humans, albeit by the help of developed software. In both cases, the technology both facilitates the analysis and improves the reliability of subsequent results and forensic evidence, but the human remains integral to the methods, and this key.

A clear future direction for forensic linguistics is to harness the power of digital humanities in the collection of relevant population data. Kredens and Coulthard (2012) outline some of the corpora that are available, but the amount of data available remains scarce. Researchers need to be afforded the time and the funds to build carefully constructed corpora of relevant population data for digital text types. Indeed, some of the datasets collected and used by stylometrists would be a good place to start, but the wider availability of population data would be a 'game-changer' in authorship analysis, both in research and casework. In a purely research context, the availability and accessibility of such corpora would allow forensic linguists to test their methods of authorship analysis in lab conditions before taking them to the police or to court. (see Solan 2013). From a soley linguistic perspective, such analysis would help in making new discoveries with regard to the composition of individual style and idiolects across text types. The benefits to casework are greater still, as such datasets could provide the all-important base rate knowledge required to bolster the reliability of stylistic approaches to authorship cases. Therefore, the creation and exploration of large-scale specialised corpora should be a priority for the field, and is an activity which lives unequivocally under the 'catch-all, big tent name' (Terras, 2016: 1637) of 'digital humanities'. Further to this, forensic linguists must continue their efforts to engage with investigators as the issue of cyber-assisted crime continues to grow. Law enforcement bodies remain in need of novel approaches to policing the online sphere, and as the work discussed above demonstrates, there is an important role for forensic linguistics to play in such endeavours.

7. FURTHER READING

Coulthard, M., Johnson, A. and Wright, D., 2017. *An Introduction to Forensic Linguistics: Language in Evidence* (2nd edn). London: Routledge.

This core textbook in forensic linguistics provides coverage of all major areas, issues and methods in the field.

8. RELATED TOPICS

Chapter 1 Spoken and Written Corpora

Chapter 10 Discourse Analysis

Chapter 12 Conversation Analysis

Chapter 14 Sociolinguistics

Chapter 18 Corpus Linguistics

Chapter 30 English Language and Social Media

9. REFERENCES

Author A. and x, 2012

Author A and x, 2016

X and Author A 2016.

Author B., 2013.

Author B., 2014.

Author B., 2017.

Abbasi, A. and Chen, H. 2005. Applying authorship analysis to extremist-group web forum messages. *IEEE Intelligent Systems* 20(5), 67–75.

Abbassi, A. and Chen, H. 2008. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems* 26(2), 1–29

Argamon, S. and Koppel, M. (2013) A systemic functional approach to automated authorship analysis. *Journal of Law and Policy* 21(2), 299–316.

Baron, N. 2013. Instant Messaging. In S. Herring, D. Stein, T. Virtanen (eds) *Pragmatics of Computer-mediated Communication.* Berlin: De Gruyter, 135-162.

Bolander, B. and Locher, M. 2014. Doing sociolinguistic research on computer-mediated data: A review of four methodological issues. *Discourse, Context & Media* 3, pp. 14-26.

Brown-Blake, C. N., and P. Chambers., 2007. The Jamaican Creole speaker in the UK justice system. *The International Journal of Speech, Language and the Law* 14(2), 269–294.

Butters, R. 2012. Retiring President's closing address: ethics, best practices, and standards. In S. Tomblin, N. MacLeod, R. Sousa-Silva and M. Coulthard (eds*) Proceedings of the International Association of Forensic Linguists Tenth Biennial Conference*. Birmingham: Centre for Forensic Linguistics, 351-361.

Cambier-Langeveld, T., 2016. Language analysis in the asylum procedure: a specification of the task in practice. *The International Journal of Speech, Language and the Law*, 23(1), 25–41.

.

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

Chaski, C. E., 2005. Who's at the keyboard? Authorship attribution in digital evidence investigations. *International Journal of Digital Evidence* 4(1), 1–14.

Cheng, E. K., 2013. Being pragmatic about forensic linguistics. *Journal of Law and Policy* 21(2), 541–550.

Cohen, William W. 2009. *Enron Email Dataset*. [online]. Available from: http://www.cs.cmu.edu/~enron/.

Corney, M., Anderson, A., Mohay, G., and de Vel, O., 2001. *Identifying the authors of suspect Email.* [online] http://eprints.qut.edu.au/8021/1/CompSecurityPaper.pdf.

Coulthard, M., 1994. On the use of corpora in the analysis of forensic texts. *The International Journal of Speech, Language and the Law (Forensic Linguistics)* 1(1), 27–43.

Coulthard, M., 2004. Author identification, idiolect, and linguistic uniqueness. *Applied Linguistics* 24(4), 431–447.

Coulthard, M., 2013. On admissible linguistic evidence. *Journal of Law and Policy* 21(2), 441–466.

Coulthard, M., Grant, T. and Kredens, K. 2010. Forensic Linguistics. In R. Wodak, B. Johnstone and P. Kerswill (eds) *Handbook of Applied Linguistics*. Thousand Oaks and London: SAGE Publications.

Coulthard, M., Johnson, A. and Wright, D.,2017. *An Introduction to Forensic Linguistics: Language in Evidence* (2nd edn). London: Routledge.

Crown Prosecution Service., 2014. *Expert Evidence*. Available from: http://www.cps.gov.uk/legal/assets/uploads/files/expert_evidence_first_edition_2014.pdf

Donath, J. 1999. Identity and deception in the virtual community in M. Smith and P. Kollock (eds). *Communities in Cyberspace*. Abingdon: Routledge, pp. 29–59.

Eckert, P. and McConnell-Ginet, S., 1998. Communities of practice: where language, gender and power all live? In J. Coates (ed.) *Language and Gender: A Reader*. Oxford, Blackwell, 484–494.

Ellison, M. and Morgan, A., 2015. Review of possible miscarriages of justice: impact of undisclosed undercover police activity on the safety of convictions. *Report to the Attorney General.*

Flowerdew, L., 2004. The argument for using English specialized corpora to understand academic and professional settings. In U. Connor and T. A. Upton (eds.): *Discourse in the Professions: Perspectives from Corpus Linguistics*. Amsterdam: John Benjamins, pp. 11–33.

Georgakopoulou, A. 2011. "On for drinkies?": Email cues of participant alignments. *Language@Internet* 8.

Grant, T., 2013. TXT 4N6: method, consistency, and distinctiveness in the analysis of SMS text messages' *Journal of Law and Policy* 21 (2), pp. 467-494.

.

Grant, T., 2017. Duppying yoots in a dog eat dog world, kmt: determining the senses of slang terms for the Courts. *Semiotica*, 216, 479–495.

Grieve, J., Nini, A., and Guo, D., 2016. Analyzing lexical emergence in Modern American English online. *English Language and Linguistics*, 21, 99–127.

Herring, S. 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, pp. 338–76.

Herring, S. 2012. Discourse in Web2.0: Familiar, Reconfigured and Emergent*. Georgetown University Round Table on Languages and Linguistics 2011: Discourse 2.0: Language and new media,* pp. 1-29.

Hoey, M., 2005. *Lexical Priming: A new theory of words and language*. London: Routledge

Howald, B. S., 2008. Authorship attribution under the rules of evidence: empirical approaches – a layperson's legal system. *The International Journal of Speech, Language and the Law* 15(2), 219–247.

Internet Watch Foundation, 2013. ChildLine and the Internet Watch Foundation form new partnership to help young people remove explicit images online. [Online]. Available at: https://www.iwf.org.uk/about-iwf/news/post/373-childline-and-the-internet-watch-foundation-form-new- partnership-to-help-young-people-remove-explicit-images-online

Jockers, M. L., and Underwood, T., 2016. Text-Mining the Humanities. In . S Schreibman, R. Siemens, and J. Unsworth (eds.): *A New Companion to the Digital Humanities*. London: Wiley Blackwell, pp. 291–306

Juola, P., 2015. The Rowling Case: A Proposed Standard Analytic Protocol for Authorship Questions. *Digital Scholarship in the Humanities*, 30 (suppl 1), i100–i113.

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

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

Koppel, M. Schler, J. and Argamon, S. 2011. Authorship attribution in the wild. *Language Resources and Evaluation* 45(1), pp. 83-94.

Koppel, M., Schler, J., and Argamon, S. (2013) Authorship attribution: What's easy and what's hard? *Journal of Law and Policy* 21(2), 317–332.

Kredens, K., 2002. Towards a corpus-based methodology of forensic authorship attribution: a comparative study of two idiolects. In Barbara Lewandowska-Tomaszczyk (ed.) *PALC'01: Practical Applications in Language Corpora*. Peter Lang: Frankfurt am Mein, 405–437.

Kredens, K. and Coulthard, M., 2012. Corpus Linguistics in authorship identification. In Peter Tiersma and Lawrence M. Solan (eds.) *The Oxford Handbook of Language and Law*. Oxford: Oxford University Press, 504–516.

McGuire, M. and Dowling, S. 2013. Cyber crime: A review of the evidence. *Home Office Research Report* 75.

McMenamin, G. R., 2002. *Forensic Linguistics: Advances in Forensic Stylistics*. Boca Raton, Florida: CRC Press.

Mosteller, F. and Wallace, D., 1964. *Inference and Disputed Authorship: The Federalist*. Reading, MA: Addison-Wesley Publishing Company Inc.

Nathan, C., in press 2016. Liability to deception and manipulation: The ethics of undercover policing. *Journal of Applied Philosophy*

Nini, A. and Grant, T., 2013. Bridging the gap between stylistic and cognitive approaches to authorship analysis using Systemic Functional Linguistics and multidimensional analysis. *The International Journal of Speech, Language and the Law* 20(2), 173–202.

Rico-Sulayes, A., 2011. Statistical authorship attribution of Mexican drug trafficking online forum posts. *The International Journal of Speech, Language and the Law* 18(1), 53–74.

Schler, J.,Koppel, M., Argamon, S., and Pennebaker, J., 2006. Effects of Age and Gender on Blogging. *Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*. Available at: http://u.cs.biu.ac.il/~schlerj/schler_springsymp06.pdf

Seargeant, P. and Tagg, C. (eds), 2014. *The language of social media: identity and community on the Internet*. London: Palgrave Macmillan.

Searle, J.S., 1969. *Speech Acts.* Cambridge: Cambridge University Press.

Searle, J. S., 1975. Indirect speech acts. In, ed. P. Cole & J. L. Morgan (eds), *Syntax and Semantics, 3: Speech Acts*. pp. 59–82. New York: Academic Press.

Silva, R., Laboreiro, G., Sarmento, L., Grant, T., Oliveira, E, Maia, B., 2011. 'twazn me!!! ;(' Automatic authorship analysis of micro-blogging messages. *Natural Language Processing and Information Systems. 16th International Conference on Applications of Natural Language to Information Systems, NLDB,* 161-168.

Solan, L., 2013. Intuition versus algorithm: The case for forensic authorship attribution. *Journal of Law and Policy* 21(2), 551–576.

Solan, L and Tiersma, P. 2004 Author Identification in American Courts. *Applied Linguistics* 25(4), 448–465.

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

Tagliamonte, S and Denis,D., 2008. Linguistic ruin? LOL. Instant messaging and teen language. *American Speech* 83 (1), pp. 3-34.

Terras, M., 2016. A Decade in Digital Humanities. *Journal of Siberian Federal University. Humanities & Social Sciences* 9(7), 1637-1650.

Törnberg, A. and Törnberg, P., 2016. Combining CDA and topic modeling: Analyzing discursive connections between Islamophobia and anti-feminism on an online forum. *Discourse & Society* 27(4), 401–422.

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.

Turell, M. T. and Gavaldà, N., 2013. Towards an index of idiolectal similitude (or distance) in authorship analysis. *Journal of Law and Policy* 21(2), 495–514.

Unsworth, J., 2002. What is Humanities Computing and What is Not? In V. G. Braungart, K. Eibl, and F. Jannidis (eds.): *Jahrbuch für Computerphilologie* 4, pp. 71–84.

NOTES

The experimental data collected with Assuming Identities Online are available via Open Access and can be found at the following link: http://dx.doi.org/10.5255/UKDA-SN-852099

Assuming Identities Online – description, development and ethical implications, was funded by the ESRC, project number ES/L003279/1.

[8805 words]