1. Editor’s Introduction: from intellectual challenges to established corpus techniques

Forensic Linguistics is the branch of applied linguistics concerned with the interface between language and law. It encompasses a wide range of topics including, but by no means limited to the language of police interviews (e.g. Heydon, 2005; Haworth, 2010) and prisoners (e.g. Mayr, 2003; Medlicott, 2001), language use in the courtroom (e.g. Conley & O’Barr, 1998; Cotterill, 2003; Heffer, 2005); legal language (e.g. Tiersma, 2000; Finegan, 2011; Goźdz-Roszkowski, 2011) and the comprehensibility of legal language such as the UK police caution (e.g. Rock, 2007). The field also covers the authenticity and credibility of language such as in the authorship analysis of a range of texts including SMS text messages (Grant, 2010), ransom demands (Shuy, 2001), suicide notes (Eagleson, 1994), terrorist manifestos (Fitzgerald, 2004) and disputed confessions (e.g. Shuy, 1998, Coulthard, 2004). Although the phrase ‘forensic linguistics’ is generally attributed to Svartvik (1968), it was only in 1994 that academics who worked at the interface of language and law were unified through the publication of a dedicated journal, Forensic Linguistics (now The International Journal of Speech, Language and Law). Prior to this, linguists working in this specialist area worked ‘in isolation from one another and were largely ignorant of the analytic procedures used by their colleagues’ (French and Coulthard, 1994: vii). Despite being a relatively young field, in the time since 1994, there has been an explosion of research articles and monographs as well as two handbooks (Coulthard and Johnson, 2010; Tiersma and Solan, 2012) with a wide range of undergraduate and postgraduate courses in forensic linguistics being taught across the world.

There are several ways to characterise the work of the forensic linguist. Cotterill (2013), for instance, makes a distinction between ‘descriptive’ and ‘investigative’ forensic linguistics. The focus of descriptive forensic linguistics is on analysing language produced at any stage throughout the legal process with a view to characterising different genres and text types. Investigative forensic linguistics, on the other hand, is concerned with analysing language that in some way constitutes a crime (e.g. ransom demands, threats, trademark infringement). A similar distinction is proposed by Coulthard and Johnson (2007) who differentiate between descriptions of the language of the law, and the linguist who takes on casework as an expert witness. However, they revised this binary distinction in their later work (2010) after acknowledging that such a ‘distinction blurred the boundary between written and spoken language’ (p. 7). They therefore characterise the work of the forensic linguist into three areas:


i) the written language of the law; ii) interaction in the legal process; and iii) linguists acting as expert witnesses (p. 7).

Published in the very first issue of *Forensic Linguistics* was Coulthard’s (1994) seminal paper which advocated the use of corpora in forensic linguistics, with particular reference to linguistic analysis used for investigative and evidential purposes. He referred to the discipline of corpus linguistics as ‘a major new resource for the forensic linguist’ (p. 27) given that in the previous 25 years, investigative case work ‘was usually undertaken as an intellectual challenge and almost always required the creation, rather than simply the application, of a method of analysis’ (p. 27). Through description and discussion of his own cases, Coulthard concluded ‘that any improved methodology must depend, to a large extent, on the setting up and analysing of corpora’ (p. 40). Over the two decades that have followed, the field of forensic linguistics has grown into a credible branch of applied linguistics, and, despite the fact that there is still considerably more work to be done (Coulthard and Johnson, 2010: 614), corpus techniques and approaches are now accepted and widely used by forensic linguists. Several prominent academics have written about the role of corpora in forensic linguistics in various outlets including handbooks (Kredens and Coulthard, 2012) and encyclopaedias (Cotterill, 2010; Cotterill, 2013). The aim of this introduction is, therefore, to bring together a disparate array of literature which utilises corpora and corpus techniques in forensic linguistics research in the form of an annotated bibliography.

Before presenting the annotated bibliography, it is necessary to describe the articles included in this Special Issue. It would be impossible to cover every aspect of forensic linguistics in a single issue, but in presenting this collection of four papers, it is hoped that a range of current approaches are represented which between them draw out some of the most interesting methodological considerations of using corpora in forensic linguistics research. The four articles selected for publication all have an applied focus; that is, the research described in these pages has potential impact at various stages of the legal process ranging from evidence gathering, to prosecution and trial, and issues of corpus construction and interrogation—including balancedness and representativeness—are importantly determined, influenced, and limited, by the data and resources available.

In the first paper, Tatiana Tkačuková argues for corpus approaches to be used in socio-legal studies. Through an analysis of judges’ use of the discourse marker ‘well’ in court cases where lay people represent themselves (litigants in person), Tkačuková highlights areas where judges can be trained to more effectively interact with litigants in person, whilst also drawing out some of the difficulties in gaining access to forensic data. In the second paper, Tammy Gales focusses on the use of grammatical stance markers in stalking threats and argues that such an analysis may provide insight into the definition and delineation between ‘threat’, ‘harassment’, and ‘defamation’, which are important distinctions in law. Gales further argues that understanding grammatical stance in stalking threats may contribute to the demonstration of a victim’s claim of feeling fear, required by law for a conviction of stalking. Gales’ research paves the way for future research which may assist threat assessors and law
enforcement agencies. In the third paper, Claire Hardaker highlights how corpus methods may facilitate forensic pragmatic analyses of impoliteness and linguistic aggression. Specifically, she identifies common responses to perceived trolling in one particular form of computer-mediated communication: posts to Usenet forums. Whilst recognising that further work is required, Hardaker’s research lays the foundation for future studies in online linguistic manipulation and online crimes such as trolling and cyberbullying. Dawn Archer and Cliff Lansley, in the final paper, argue that whilst making quick decisions about deception in high-stakes, real time scenarios is less than desirable from an academic perspective, practitioners are required to do this, so they question how technology can assist them to do their job better. They explore the role that corpus techniques, particularly the use of Wmatrix, may play in determining the validity of language features previously identified as being related to deception detection in high stakes situations.

2. **Annotated Bibliography of the use of Corpora and Corpus Techniques in Forensic Linguistics Research**

In bringing together this annotated bibliography of over 50 references, it is hoped that the development of corpus linguistics in forensic linguistics, as well as the multitude of ways in which corpora have been developed and used in a variety of different applications, will be shown. The list is necessarily restricted and whilst it is representative of the vast array of data and methods used, it should not be considered exhaustive. Entries have been listed chronologically for convenience. All entries have been classified into broad themes using Coulthard and Johnson’s (2010) tripartite distinction, that is as being most relevant to: A) the written language of the law; B) interaction in the legal process; or C) linguists acting as expert witnesses, or rather research related to using language as evidence. It is perhaps also worth noting that whilst the bibliography contains research by established and eminent forensic linguists, many of the entries are written by academics who would not normally consider themselves forensic linguists. Their work, nonetheless, either draws on forensic data, or has implications for the field, which demonstrates how forensic linguists draw on a multitude of linguistics sub-disciplines and specialities, rather than being confined to only research which can clearly be labelled as forensic. What becomes apparent from the annotated bibliography is that although there are some notable exceptions of corpus-driven approaches, largely confined to exploring linguistic patterns and constructions in legal discourse, forensic linguistics research which utilises corpus linguistics has, to date, largely utilised corpus-based approaches. In their handbook, Coulthard and Johnson (2010) highlight the relationship between power and (dis-)advantage in forensic linguistics, and predict critical forensic linguistics—the analysis of power and (dis-)advantage relationships in legal texts—as an area that will be taken up in the next two decades (p. 602-3). Given the increased reliance on corpora and corpus analysis, it seems likely too that Corpus-assisted Discourse Analysis and Corpus-Based Critical Discourse Analysis will gain prominence, particularly in research which explores interaction in the legal process.


<table>
<thead>
<tr>
<th><strong>Reference</strong></th>
<th><strong>Annotation</strong></th>
<th><strong>Theme</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Winter, E. 1996. ‘The statistics of analysing very short texts in a criminal context’, in H. Kniffka, S. Blackwell &amp; M. Coulthard (eds) <em>Recent Developments in Forensic Linguistics</em>, pp. 141-180. Frankfurt am Main: Peter Lang.</td>
<td>Acknowledging that forensic texts are typically very short, this vocabulary analysis research analyses three criminal statements of between 481 and 616 words to determine whether authorship can be ascertained.</td>
<td>C</td>
</tr>
</tbody>
</table>
This is a post-refereed version of an article which has been published by Edinburgh University Press:


<table>
<thead>
<tr>
<th>Authors (Year)</th>
<th>Title</th>
<th>Pages</th>
<th>Notes</th>
</tr>
</thead>
</table>


Using a corpus of Dutch university writings, measures linguistic distances between texts for authorship purposes.

Outlines a sampling and testing method in an attempt to quantify results from authorship analysis investigations. Tests the method against a corpus of authors' texts which decreases in size.

Uses statistics to discriminate between a corpus of 33 authentic and 33 simulated suicide notes.

Focuses on the construction of argumentation in a corpus-based genre analysis of judicial texts.

Analyses legal texts for linguistic and pragmatic functions, focussing in particular on tense, aspect, modality and verbal constructions.

Corpus-based investigation into lexical and syntactic features of idiolectal variation, based on the speech of five White House Press secretaries.

Analyses a corpus of witness examination transcripts from famous Public Inquiries in Northern Ireland, England and Scotland to explore how witness examination has developed as argumentative dialogue.

Explores interruptions in a corpus of transcripts from four Chinese criminal trials.
This is a post-refereed version of an article which has been published by Edinburgh University Press:


A corpus-based study into the textual organisation patterns, particularly conjunctions, in a 200,000 English word corpus of European Union legal and business documents. Results are compared to the BNC.


Describes a corpus of criminal statements, police interrogations and civil testimony that was annotated for language-based cues to deception and to verify details relevant to the cases. Article focuses on methods for building corpora for deception research and pertinent issues.


Through an analysis of a corpus of US Supreme Court judgements, the verbal and adjectival tools, and underlying patterns of judges' argumentative positions are explored.


More generally explores useful variables for analysis in texts produced by different authors based on an analysis of 381 different measures. Conclusions are made which draw out links between this work and forensic authorship analysis.


Outlines two quantitative approaches to authorship analysis which involve comparing forensic texts with a general corpus in order to determine markedness and saliency.


Discusses how computational analysis may be useful in the comparison of documents for authorship, particularly plagiarism, with a focus on the reliable handling large quantities of data.
This is a post-refereed version of an article which has been published by Edinburgh University Press:


Describes the 8 million word Legal Māori Corpus, which consists of printed legal texts from the 1920s to present day. Focusses in particular on the word mana.


Using four 500,000 word corpora containing commercial law texts from academic journals, case law, legislation and legal documents, six adjective/adverb sets which are prominent in legal discourse are analysed to explore how legal writers communicate meaning. Results are compared to the BNC and British Academic Written English Corpus.


A genre analysis of three one million word corpora (U.S. patents, contracts, and regulations on technical matters), which are computationally analysed for features of vagueness.


Focusses on adverbial expressions of attitude and emphasis in United States Supreme Court decisions.


Adopts a corpus-based approach to the analysis of variation in legal English with reference to recurrent linguistic patterns.


A corpus of Polish Suicide Notes is described along with its annotation system which codes for document structure, textual content and linguistic features.
This is a post-refereed version of an article which has been published by Edinburgh University Press:


Wmatrix is used to calculate frequency counts, key words, and concordance analysis of 250 extremist terrorist statements to gain insight into motivations for terrorism. Concludes by discussing how the results may feed into counter-terrorism strategies.

Investigates the occurrence of lexical bundles in four legal corpora (academic law, case law, legislation, and legal documents) to show differences between lexical bundle types and their functions.

In controlled tests of authorship, several corpora consisting of English, German, Polish, Ancient Greek and Latin prose texts were ‘damaged’ to investigate the link between a ‘dirty’ corpus and attribution accuracy.

Argues for the role that corpus methods can play in legal interpretation, particularly how plain or ordinary meanings of terms in given contexts i.e. legal ambiguity, can be quantified.

Corpus-based study into attribution and judicial control in appellate judgements in mainland China, Taiwan and Hong Kong courtrooms.

Adopts a corpus-based approach to investigate whether watching foreign, discipline-specific language television programmes may lead to the learning of technical legal vocabulary.

A corpus analysis of the term "human rights" is carried out on a corpus of American and British print media data.


A

B

C

A

C

A

Describes a corpus of 100 personal narratives, authored by 20 writers, used to determine whether formulaic sequence usage is sufficient to differentiate between authors.


Assesses whether the web as corpus is reliable as an evidential tool in forensic authorship analysis, with specific reference to idiolectal co-selections.


Analyses modality in English translations of Chinese legislation to highlight problems in translation. Data is compared to a parallel corpus of China's Legal Documents.


Adopts a corpus-based approach to explore the issue of conjunctive cohesion in legal documents.


Uses Wmatrix to differentiate between language used by 78 genuine and deceptive pleaders during televised press conferences.

### References


