Corpus Linguistics and Big Data: Researching language in historical trial records of the 18th and 19th century .

Dr Alison Johnson, University of Leeds, a.j.johnson@leeds.ac.uk

The big data hype has hit business and academia alike, but what is big data? For linguists the 650-million-word Bank of English or the 100-million-word BNC are huge, but in the electronic world big data means terabytes and petabytes, not gigabytes of data. So, as corpus linguists our big data (BNC, Bank of English) are really minute, as are the small, specialised corpora (Cameron and Deignan, 2003; Flowerdew, 2004) that we work with in forensic linguistic research. However, through the analyses of these large collections of specialised textual data, we can develop tools, methods and linguistic findings that can be of benefit to those developing tools for text mining big data.

Drawing on a wealth of corpus-based forensic research (e.g. Archer, 2005; Cotterill, 2003, 2010; Coulthard, 1994; Finegan 2011; Heffer, 2005; Kredens 2002; Wright 2013), I illustrate my recent corpus-based forensic research (e.g. Johnson and Clifford 2011; Johnson 2014; Johnson and Wright 2014) from work with historical corpora. I show how corpus and computational methods and tools can reveal differing lexical representations of prosecution and defence views of offences, offenders and victims in 18th century rape trials and 19th century trials involving insanity defences. Research reveals the historical underpinnings of many of today's rape myths and the power of prosecution and defence barristers to produce competing narrative biases for the jury to evaulate.

Conclusions point to the continuing and future importance of corpus-based and corpus-driven research and software development, as forensic linguists respond to the growing amount of data available for study and analysis.