Data and Ethics in Forensic Linguistics

Kate Haworth
Aston Institute for Forensic Linguistics
Centre for Research on Spoken Interaction in Legal Contexts
Birmington, UK

Abstract

In this paper I will focus on an aspect which is at the heart of ethical conduct in forensic linguistics, namely data. Our data can take various forms, but most commonly involve highly sensitive materials generated not by the researcher, but as a naturally occurring part of the context we seek to research. This gives rise to particular ethical challenges and dilemmas which are perhaps less prominent in other fields. Examples include recordings and transcripts of police interviews, courtroom judgments, and recordings of calls to the emergency services. It can also include casework materials provided to academics being instructed as expert witnesses. Gaining access to such data is often a huge challenge in itself, which requires careful handling to build trust and mutual understanding. Once access has been granted, a new set of challenges arises, particularly around questions of confidentiality, data protection, and data ownership. As an alternative to having to overcome such high barriers to access, researchers are increasingly sourcing forensically relevant data online, including using open-access, high-visibility sites such as YouTube. However, the fact that data are publicly available does not automatically mean that there are no ethical issues with using them in our research; something which perhaps needs wider recognition and reflection. The stakes in the FL context are high, and so it is crucial that, as a collective community of researchers, we start to address these responsibilities around our data more cohesively and explicitly, rather than relying on the ad hoc practice of individual researchers. There is also a dearth of guidance or instruction to help newcomers to the field to navigate these strange waters. If done appropriately, the FL community has the opportunity – and arguably the responsibility – to set the gold standard for data ethics in linguistics and beyond. By the same token, the risk of getting it wrong deserves to be taken more seriously: if not handled well, we risk casting a shadow over the reputation of the field, discrediting the work of linguistic experts in the eyes of the courts and other professionals, as well as damaging co-operative relationships we are so carefully developing with data providers. In this paper I will focus in particular on anonymisation as an academic practice, including considering the competing demands of protecting the confidentiality of individuals, while preserving the data as closely as possible to the original for research integrity and authenticity. I will also discuss the practical challenges of anonymising linguistic data, including audio and video data, considering questions such as what counts as ‘identifying’, and evaluating different techniques and strategies available to us, some of which risk giving rise to serious unintended and potentially unethical side effects, especially when handling features such as gender and ethnicity.