Abs-5

Tony McEnery and Andrew Hardie (Lancaster University)

Research ethics in corpus linguistics

While research ethics are as critical for corpus linguistics as for any other branch of linguistics, relatively little consideration has been paid in the literature to ethical issues in corpus construction and exploitation. Although some authors have directly considered their work in relation to ethical issues, for example Hasund (1998), Sampson (2000) and Rock (2001), the central textbooks in the field, including Sinclair (1991), Kennedy (1998), Biber et al. (1998), and McEnery and Wilson (2001), do not treat ethical issues in any depth. This may be because corpus linguists have in many cases 'inherited' their ethical good practices from guidelines developed, for example, for applied linguistics in general. For example, the British Association of Applied Linguistics has a well-developed set of ethical guidelines which are clearly relevant to corpus builders (see http://www.baal.org.uk/dox/goodpractice_full.pdf).

We will argue, however, that there are questions specific to corpus linguistics which may not be fully addressed by guidelines from outside the field, and thus that research ethics is an area that corpus linguistics should consider in more detail. There are four main groups of such questions. Firstly, in collecting a spoken corpus there are ethical issues relating to the respondents. These relate primarily to privacy – not only of the respondent, but also of the people they are recorded speaking to, and moreover of the people they are recorded talking about. A second set of ethical issues must be addressed in the process of construction of a written corpus. In particular, what is the appropriate attitude to take towards potentially offensive, immoral, or illegal textual data? A third group of questions relate to the sometimes vexed question of the distribution of corpus data. To what extent are corpus distributors ethically obliged to consider whether the purposes to which the data will be put would be approved of by the original donors/collectors of the data? Finally, there are issues that must be faced by any user of corpus data – in particular, the ethical imperatives to take all steps to make sure their analysis is replicable, and to record and preserve aspects of the research method that underlie, but are not contained within, their published results.

While the corpus linguistic literature is mostly silent on ethical issues, it does generally embody good ethical practice. There are, however, a number of exceptions – instances of relatively poor practice in published corpus research. Some occurred during the infancy of corpus linguistics as a (sub)discipline, but some are more recent. We will review some examples of such poor practice, and suggest that, as a corrective to these relatively prominent bad examples, it is high time that more explicit regard is given to issues of research ethics in corpus linguistics.

References

Biber, D, Conrad, S and Reppen, R (1998) Corpus linguistics: investigating language structure and use. Cambridge: CUP.

Hasund, K. (1998) 'Protecting the innocent: the issue of informants' anonymity in the COLT corpus', in A. Renouf (ed.) Explorations in Corpus Linguistics, Rodopi, Amsterdam, pp 13-28.

Kennedy, G. (1998) An Introduction to Corpus Linguistics. Harlow: Longman.

McEnery, T. and Wilson, A. (2001) Corpus Linguistics (2nd edition). Edinburgh: EUP.

Rock, F. (2001) 'Policy and Practice in the Anonymization of Linguistic Data', International Journal of Corpus Linguistics, Volume 6, Number 1, pp 1-26.

Sampson, G.R. (2000) CHRISTINE Corpus, Stage I: Documentation. Available at www.grsampson.net/ChrisDoc.html

Sinclair, J. (1991) Corpus, Concordance, Collocation. Oxford: OUP.