The web, laws and ethics

Tony McEnery
Introduction

- Practical issues – some are familiar
- Some are rarely touched upon
- Other are almost never discussed at any length
- This talk surveys legal issues in web based corpus collection and ethical issues in general.
- I speak from a Western standpoint and from my own experiences with research projects
Law and the Web

• The fundamental issue – do you have the right to gather and distribute?
• The World Wide Web – an opportunity
• The World Wide Web – a challenge
• BootCat (Baroni and Bernardini, 2004) – a system for collecting web corpora
• ‘Web as Corpus’
• Chat room data (Claridge 2007, Thelwall 2008, King 2009)
Diverging Views

• The web as ‘a fabulous linguist’s playground’ (Kilgarriff and Grefenstette 2003: 333).

• The web ‘can in no way be considered a representative sample of language use in general’ (Leech 2007: 145). A

• Although the web *can* be useful, the ‘more sophisticated needs of the working linguist may be better fulfilled by means of traditional corpora’ (Lew 2009: 298).
An Issue - Genre

• How does one determine the genre of a web document?
• Baroni and Bernardini (2004: 1315) suggests that one in three of the webpages recovered may not be in the desired genre.
An Issue - Law

• Copyright laws apply to documents available on the web exactly as they do to print documents
• Financial loss as a measure
• Corpus data needs to be widely available if corpus linguistics is to be replicable
• This is an ethical imperative for the researcher.
Ways of Addressing Copyright

• Treat text from the web the same as any other text (Baker et al. 2004)

• Collect data only from sites which explicitly allow the re-use and redistribution of text (e.g. Wikipedia)

• Collect data without any regard to seeking permission and make it available to other researchers through a tool that does not allow copyright to be breached (Davies, 2010)

• Fair use?

• Redistribute a list of the web addresses from which the corpus has been collected, not the corpus itself
Ethics

- Somewhat neglected in corpus linguistics
- A problem solved by others, who produce guidelines we can use?
- [http://www.baal.org.uk/about_goodpractice_full.pdf](http://www.baal.org.uk/about_goodpractice_full.pdf)
Ethics and Respondents

• Example: spoken BNC
• Issue: the sacrifice of privacy
• Issue: securing sensitive information
• The BNC did address such issues, but ...
• What about the tapes? More specifically, what about those who were recorded coincidentally?

... comment that one of their schoolmates, identified by Christian name, behaves like a whore. This person is entitled to anonymity as much as the speakers, and arguably more so: she signed no release form for the corpus compilers. When well-known public figures or institutions are mentioned, the BNC compilers seem to have felt that there was no need to anonymise the references at all. Clearly, if someone announces that he has just bought the latest album by a named pop singer, there is no point in concealing the singer’s name. But it depends on what is said. One of the CHRISTINE texts contains a series of quite damaging remarks about the management of a secondary school, named in the BNC file. In another case, speakers comment adversely on the sexual morality of a named American actress. Even American actresses, surely, are entitled to have their honour guarded by corpus linguists.

• Yes, but how do we help?
During nineteen ninety one the Board has been delighted to open new areas of work in Inverness where our first designated place and associated hostel was opened on a most happened-- happy day by Sir Russell <gap desc="name" reason="anonymization"> .

In Elderslie near Paisley <pause> where Lady <gap desc="name" reason="anonymization"> the wife of last year's Lord High Commissioner opened our fourth senile dementia unit.
• Leech and Weisser (2003) contained credit card details that had to be anonymised

• The Lancaster Corpus of Children’s Writing (Smith et al. 1998) corpus contained the personal details of young children

• file HE7 in the BNC
A 275
Well don't you think that it's really rather improper for you to be doing this?

A 276
After all people are entitled to some secrecy aren't they, about their <unclear> breaking furniture> You don't feel that there's any need at all to give any explanation of your behaviour?

A 277
You don't think that an explanation is due here?

A 278
This information after all should have received confidential and does belong to other people, doesn't it?

B 279
What I thinks embarrassing is that you're just stealing it so that you can make money aren't you?

A 280
And you're just stealing it you're just stealing it so that you can make money aren't you?

A 281
Do people have a right to have their health records confidential do they not? 

A 282
Have you got nothing to say what so ever?

B 283
'Fraid not, no. <event: "footsteps"> 

A 284
Robert is not alone in selling personal information from data banks.
• Issue: Video and Audio
• ‘[a]nonymisation is extremely difficult, if not impossible, because the data is so rich’ (Reiter 2004: 2).
• Hasund (1998: 16-17):

In the invitation to take part in the research project ... the following promise was given to the COLT recruits: ‘You and the people you have recorded are guaranteed full anonymity’. There were lengthy discussions among the researchers working on the corpus of what was implied by the term ‘full anonymity’, resulting in an agreement to delete all surnames and addresses in the transcription, but leave all first names unchanged. Considering that the recordings were made in a huge city like London, and the recruits were pupils and not public persons connected to specific positions at specific universities of companies, this level of anonymization was considered sufficient for the protection of personal identities.
Ethics and Corpus Builders

• EMILLE corpus (Baker et al. 2004)
• Opportunistic corpus
• A religious organization saw the opportunity to contribute to our corpus as a way of distributing their material and thus gaining converts
• **Who it is alright to hate**
• A question of research ethics – ethics relating to the conduct of scientific experiments
• The underlying problem is embedded in one of the great strengths of the corpus approach: corpora are multifunctional
Discusión

- Should a corpus be censored? For example, the BNC has been used to explore swearing in English. This is possible because a choice was made when building the corpus not to censor the data. How defensible is that decision? Are there circumstances in which you would consider censoring corpus data? If so, on what grounds would you do so?
Ethics and Corpus Distributors

• Issue: collecting data from outlawed groups
• Issue: who is funding you and why?
• Issue: may the corpus become illegal in certain jurisdictions if certain decisions are made?
• Issue: keeping data intact and available.
Ethics and Corpus Users

- Issue: what may your corpus analysis cause to happen?
- Issue: making your analyses available to future researchers
- Issue: making your tools available to future researchers
- Automatic and manual analyses have differing issues, but the imperative to preserve the analysis is constant
- Issue: How will others interpret your results? What may be the impact of an interpretation you did not expect?
Ethically Problematic Research – Some Case Studies

• Corpus builders do not routinely produce corpora that are overtly unethical.
• Nonetheless, it is possible to find instances of poor practice.
• The BNC spoken corpus has a somewhat haphazard approach to anonymisation.
• The Survey of English Usage – serruptitious recordings
• Lost data and methods – Philips (1989)
• The most problematic ethical choices are firmly in the past
• Though corpus linguistics has made mistakes in the past, it has learnt from them
• Prediction: the interaction between corpus linguistics and legislation will intensify
Discussion

• Imagine you are building a spoken corpus, and you are collecting data by audio-recording spontaneous conversation. Given that you are following standard ethics procedures to get informed consent from all speakers, what steps could you take, in designing the data collection and transcription procedure, to minimise the observer effect that will inevitably result from following these procedures?
Homework!

- The next time you find yourself in conversation with a group of friends, imagine that you are secretly recording the conversation. Make a mental note (and, later, a written note) of the ethical issues that might arise if you were intending to transcribe and then publish the data without the participants’ knowledge or consent. Repeat this ‘experiment’ in a number of other contexts. What common issues emerge? Do any ethical issues seem bound to certain types of interaction?
Further Reading

• McEnery & Hardie (2012)
• Hundt et al. (2007)
• Leech (2007)
• Baroni et al. (2008)
• Hoffmann (2007a, 2007b)
• McEnery et al. (2006: 77-9)
• Rock (2001)
• Hassund (1998).
Bibliography


