Ethics considerations for Corpus Linguistic studies using internet resources

Ansgar Koene
University of Nottingham
Ansgar.koene@nottingham.ac.uk

Svenja Adolphs
University of Nottingham
Svenja.adolphs@nottingham.ac.uk

1 Introduction

With the rising popularity of public and semi-public communication channels such as Blogs (late 1990s), Wikipedia (launched in 2001), Facebook (launched in 2004), Reddit (from 2005) and Twitter (from 2006), the Internet has become an increasingly fertile medium through which to collect substantial data sets of written language. Additional features that make online communication platforms attractive include the comparatively low effort and cost associated with data collection and the unobtrusive nature of the collection process, which can often be performed 'behind the scenes' using application programme interfaces (APIs) or web scraping techniques, depending upon the affordances of the specific type of social media studies (e.g. Twitter, Blogs). While the unobtrusive nature of the methods offers the advantage of ensuring that observed conversations are not unduly influenced by the researcher, it raises ethical concerns around issues of privacy violation, informed consent and the right to withdraw.

In this paper we will discuss some of the ethical concerns around the use of online communications data. We will start by looking at the current guidelines by the British Association for Applied Linguistics (BAAL). Next we will discuss some of the core difficulties related to identifying 'publicness' of Internet-based information. This will lead to a discussion about ethical responsibilities when dealing with 'public' online communications, and how this issue is being addressed in current corpus linguistics research.

2 BAAL guidelines

In its discussion of Internet research (section 2.9) the "Recommendations on Good Practice in Applied Linguistics" guidelines (BAAL, 2006) starts by acknowledging that it is often difficult to establish if a specific online communication should be considered to be in the private or the public domain. The distinction between private and public domains, however, has significant consequences for the nature of consent and confidentiality, and how they are subsequently addressed when dealing with the data. In this respect, the BAAL (2006) guidelines advise:

> In the case of an open-access site, where contributions are publicly archived, and informants might reasonably be expected to regard their contributions as public, individual consent may not be required. In other cases it normally would be required. (BAAL, 2006)

This guideline is often interpreted to mean that when handling online data that could be reasonably considered as public communications, it is not necessary to notify participants about the act of data collection, or the analysis and publications that are based on their data. The nature of online data collection, however, is such that unless explicitly informed, participants will otherwise have no way of knowing that their communications are being observed for research purposes. This type of data collection might, therefore, also be considered as 'covert research', for which section 2.5 of the BAAL (2006) guidelines states that:

> Observation in public places is a particularly problematic issue. If observations or recordings are made of the public at large, it is not possible to gain informed consent from everyone. *However, post-hoc consent should be negotiated if the researcher is challenged by a member of the public*. (BAAL, 2006) [emphasis added by us]

The final sentence is especially important, and problematic in the context of Internet-mediated research, since the cover nature of the data collection means that participants are effectively denied this opportunity, unless the researchers make an explicit effort to inform about their actions.

Section 2.5 of the BAAL (2006) guidelines concludes with the statement that:

> A useful criterion by which to judge the acceptability of research is to anticipate or *elicit, post hoc, the reaction of informants when they are told about the precise objectives of the study*. If anger or other strong reactions are likely or expressed, then such data collection is inappropriate. (BAAL, 2006) [emphasis added by us]

In the context of internet-mediated research, this implies that as a minimum requirement researchers should, at the end of the data collection period, post a message about the research on the communication platform, offering some form of 'data withdrawal' procedure for any participant who wishes to make use of it. In essence, such an approach emphasises that process of 'opting-out' rather than 'opting-in'.

3 Public – Private distinction

Distinguishing between public and private

communications online is probably one of the most contentious issues when trying to implement the current guidelines on Internet-mediated research ethics. Bruckman (2002) provided the following criteria for deciding if online information should be considered as 'public', and therefore, "freely quoted and analyzed […] without consent".

- It is officially, publicly archived
- No password is required for archive access
- No site policy prohibits it
- The topic is not highly sensitive.

Unfortunately, a number of potential problems persist with this sort of criteria. In a digital era characterised by Google-caching, retweeting, 'Like' buttons and other means of information replication and proliferation, what is the true meaning of "officially, publically archived"? Furthermore, in an age of 'big data' practically every communication is automatically archived by default, with publically accessible data archives generated without the user ever needing to formulate a conscious decision about the process. Blogging software, for instance, will frequently default to a mode where past blog posts are archived in a publically accessible format.

Social media, such as Twitter, introduce further problems for the Public-Private distinction. Even though it was built as a 'public broadcast' platform which people are generally aware of, it is nevertheless often used as a means for communication within networks of friends, with little intention of broadcasting content to a wider audience. In such instances, social interaction upon Twitter might be viewed more like a private conversation in a public space rather than radio broadcasts.

4 Responsibilities when dealing with public communication

A core rationale underpinning the concept that use of data from 'public' forums, such as Twitter, does not require consent is based on the premise that users of the forum have in effect already consented when they accepted the 'term and conditions' of the forum. The current reality of Internet usage, however, is that the 'terms and conditions' policies of Internet sites are rarely read and are generally formulated in ways that are too vague and incomprehensible to constitute a means of gaining true *informed* consent (Luger, 2013).

Even if users of a public forum are comfortable with the idea that their conversations may be seen and read by people they are not aware of, this does not necessarily imply that they would also be comfortable with having a large corpus of their communications analysed with the potential of generating personality profiles that intrude further into the privacy of the individual than any of their individual messages (Kosinski, Stillwell, and Graepel, 2013). This point is particularly relevant in the current climate of social media analytics where stories of unethical behaviour for commercial or security related gain are flooding the mainstream media. It is here that academia has a responsibility to enter into the discussion of what constitutes good, ethical conduct. This may be achieved by being transparent about the goals and methods of the research and gaining true *informed* consent from participants, or at least providing them with the option to withdraw.

5 Privacy vs. The greater good

So far we have discussed the issue of Ethics of Internet-based research primarily from the perspective of 'respect for the autonomy and dignity of persons', e.g. privacy. Other important ethics considerations, concerning 'scientific value', 'social responsibility', and 'maximizing benefits and minimising harm' must also be taken into consideration (BPS, 2013). When considering and conducting research studies related to preventing of socially unacceptable behaviours, such as bullying for instance, not all parties in the conversation data-set are necessarily equal. While seeking consent from the target of the bullying behaviour would be ethically required, asking consent from those who perform the bullying may not take priority in every context.

6 Conclusion

When considering the ethics of Corpus Linguistics studies using Internet-mediated resources we conclude that the concept of a binary divide between public and private communication is fundamentally flawed. Furthermore, we argue that the idea that the 'public' nature of a communication platform provides a carte-blanch for accessing the data hosted on it is highly problematic, creating a key issue for corpus linguists who analyse different types of discourse on the internet.

Acknowledgements

This work forms part of the CaSMa project at the University of Nottinghan, HORIZON Digital Economy Research institute, supported by ESRC grant ES/M00161X/1. For more information about the CaSMa project, see [1].

[1] 1 http://casma.wp.horizon.ac.uk/

References

British Association for Applied Linguistics, 2006, *Recommendations on Good Practice in Applied Linguistics*. Available online at http://www.baal.org.uk/dox/goodpractice_full.pdf

British Psychological Society, 2013. *Ethics Guidelines for Internet-mediated Research.* INF206/1.2013. Leicester.

Bruckman, A., 2002. *Ethical Guidelines for Research Online*. http://www.cc.gatech.edu/~asb/ethics/

Luger, E., 2013. *Consent for all: Revealing the hidden complexity of terms and conditions.* Proceedings of the SIGCHI conference on Human factors in computing systems, 2687-2696.

Kosinski, M., Stillwell, D. and Graepel, T., 2013. *Private traits and attirbutes are predictable from digital record of human behavior.* PNAS 110 (15): 5802-5805.