Methodology of the research project

The research project was funded under the British Academy/Leverhulme Small Grants scheme.

The below is a section of the methodology paper. The full paper will be available here once it is published.

Methodology

The study is based on a mixed methods approach. The first stage involves a linguistic analysis of social media and forum platforms using corpus linguistics tools for semi-automated text analysis (see below). The focus on lexical choices the participants make in their posts allows for a comparative analysis between the main themes discussed by LIPs, MFs and other users. The second phase of the project draws on the results of the corpus linguistics analysis to create a coding system for the content analysis which investigates the quality of advice and information provided by MFs. This article focuses on the first phase in order to present an overview of LIP – MF online interactions as well as quantify LIPs' main concerns, MFs' roles and other users' views. The article also presents an opportunity to illustrate how corpus linguistics methods can help address socio-legal research questions.

A corpus is an electronic collection of texts. What makes corpora distinct from many other electronic collections is that corpora are built with a set of principles in mind designed to improve the robustness of the subsequent research (our approach is described below). A corpus approach allows researchers to extract individual lexical features from texts, quantify them and then put them into the wider context by identifying regular patterns[1]. Several analysis methods and software packages have been developed for corpus linguistic research which offer algorithms for identifying such features and patterns. Traditionally, corpora were compiled for lexicographical purposes, e.g. compiling dictionaries, so the size of corpora needed to be large enough to extract multiple uses of individual words and represent a wide range of genres. Small specialised corpora, such as the one built for the purposes of this study, allow for a more detailed analysis of a smaller number of whole texts and usually contain detailed metadata about users or types of texts[2]; for instance, the analysed corpus has in-built metadata about the type of online source and type and role of user. As a data-driven bottom-up approach, corpus linguistics provides an objective basis for the analysis and eliminates weaknesses that are sometimes associated with such qualitative methods as discourse analysis or content analysis (e.g. subjective conclusions, selective extraction of features or de-contextualised examples – see McEnery and Wilson[3]).

---

[1] S. Thornbury 'What can a corpus tell us about discourse?' In O'Keeffe, Anne, and Michael McCarthy, eds. *The Routledge handbook of corpus linguistics* (2010) 276.

[2] T. McEnery, A. McEnery, R. Xiao and Y. Tono, *Corpus-based language studies: An advanced resource book*. (2006).

[3] A. M. McEnery and A. Wilson *Corpus linguistics: an introduction* (2001), at 115.

The corpus built for this study is a small specialised corpus of 178,811 words (70,196 words in the LIPs' sub-corpus; 39,417 in the MFs' sub-corpus; 69,194 words in the other users' sub-corpus). Designing a corpus is a key part of the objectivity of the subsequent analyses; the objectivity is ensured by a careful consideration of the following three principles: sampling, balancing and representativeness of the texts (Nelson[4]; McEnery et al.[5]). Balancing and representativeness are primary considerations for smaller specialised corpora. The representativeness of a corpus is determined by the situational context and the linguistic context[6]. The representativeness of the situational variability within this specialised corpus was established via the above-mentioned criteria for choosing social media groups and online forums (i.e. regular use, appearance in top searches, wide range of forums/social media groups, representation of both genders). The representativeness of the linguistic context was ensured by the fact that all threads include all the posts by LIPs, MFs and other users and the overall context of the thread is retained.

The principles of sampling and balancing rely on selecting adequate texts representative of a wide range of situations typical for the communicative context[7]. The sampling of the corpus was largely determined by the availability of relevant queries within the chosen period. The criterion of balancing is met by incorporating a wide range of threads into the data sample. Despite the fact that the methodological approach is based on a predominantly quantitative method of corpus linguistics, the research team decided to manually choose relevant threads within the delineated period instead of a data scraping approach because many threads on forums and Facebook groups included links to different documents, judgements, newspaper articles or even jokes. Choosing the threads manually meant that only relevant threads were included and there was more control over which threads and LIPs' queries would be incorporated to ensure a wide range of queries and a variety of MFs responding to them. Since balancing is dependant on the variety of threads, the number of threads, wide range of posters within threads, and the size of the sub-corpora, the careful choice of threads strengthened the corpus. Given the criteria chosen for sampling, i.e. a proportionate number of relevant threads within the time frame, balancing of the corpus was limited by the number of relevant queries available in individual forums or Facebook groups.

The design of the corpus is equally important. It is common to segment a corpus into sub-corpora that represent, for example, different elements of the study or different types of text. To fulfil the aims of the project and not eliminate the differences between individual user groups or social media platforms, it was important to create a flexible system of sub-corpora which could be easily combined for different types of comparative analyses. Each contribution in a thread was therefore assigned the role of poster (original poster, MF, poster), the type of source (name of Facebook group/public forum) and the type of MF (individual MFs were tagged with additional information where available, e.g. ex-LIP, ex-solicitor, forum moderator). As a result, the sub-corpora created can be combined in a

[4] M. Nelson, 'Building a Written Corpus: What are the basics' In O'Keeffe, Anne, and Michael McCarthy, eds. *The Routledge handbook of corpus linguistics* (2010) 56.
[5] McEnery et al., op. cit., n. 26, pp. 13-21.
[6] A. Koester, 'Building Small Specialised Corpora' In O'Keeffe, Anne, and Michael McCarthy, eds. *The Routledge handbook of corpus linguistics* (2010), 69–71.
[7] McEnery et al., op. cit., n. 26, p. 125.

flexible way according to the role of the posters and the type of forums/group to enable a range of comparative analyses.