Chapter 2

Corpus Linguistics and Ethics

Gavin Brookes and Tony McEnery

Lancaster University

Abstract

In this chapter, we explore the ethical considerations attending to research and practice in corpus linguistics. Despite the ubiquity of ethical dilemmas in corpus construction and use, there has been scant literature dedicated to ethical practices within the discipline. This gap is particularly pronounced given the increasing engagement with digital and online data sources, which pose unique ethical challenges regarding issues such as consent, privacy, and the public-private dichotomy. The chapter addresses these ethical considerations, and more besides, from the inter-related perspectives of research participants, corpus builders, distributors, and users. Importantly, the chapter highlights how ethical considerations are not confined to discrete stages of corpus linguistic projects but, rather, are interwoven throughout the research lifecycle. Key issues addressed include informed consent, participant anonymity, the ethical implications of using publicly available versus private communications, and the responsibilities of corpus users to ensure the meaningful, truthful, and fair representation of their findings. The chapter aims to respond to the need for more nuanced ethical guidelines that reflect the diversity of data sources and research contexts that characterise contemporary corpus linguistics, advocating for a reflective, case-by-case approach to ethical decision-making.

Introduction

This chapter explores ethical considerations underpinning research and practice in corpus linguistics. Corpus linguistics is, broadly, an umbrella term which denotes a collection of methods but also a field of linguistics research concerned with analysing language based on large, representative bodies of naturally occurring language in use (for introductions, see McEnery and Wilson 2001; McEnery and Hardie 2012; Brookes and McEnery 2020). While most corpus linguists are likely, at some point or another, to have encountered ethical considerations in relation to how they design, build and use corpora, surprisingly little has been written about ethical practice in corpus linguistics. As McEnery and Hardie (2012, p. 60) observe, '[w]hile the legal issues involved in corpus construction have been considered widely, less consideration has been paid in the literature to ethical issues in corpus construction'. While we can increasingly encounter ethical considerations being discussed in informal contexts, such as in exchanges between corpus linguists on social media and within corpus linguistics mailing lists, most of the major introductory works in corpus linguistics (e.g. McEnery and Wilson 2001; Biber et al. 1998; Hunston 2022) barely scratch the surface in terms of discussing ethical considerations, meanwhile chapters dedicated to discussing ethical considerations are absent from major corpus linguistics handbooks (e.g. Biber and Reppen 2015; O'Keeffe and McCarthy 2021).

This trend reflects the broader lack of explicit treatment of research ethics in the Applied Linguistics literature (at least as compared to the amount of methodologically oriented

publications published each year – see Sterling and De Costa (2018, p. 166)). With regard to corpus linguistics specifically, McEnery and Hardie (2012, p. 60-61) suggest that this lack of detailed attention to ethical considerations may be because corpus linguists tend to utilise existing ethical guidelines for gathering and using data already developed by professional linguists. For example, the British Association for Applied Linguistics has a well-developed set of ethical guidelines which are clearly relevant to corpus builders. In addition, most universities and other research organisations have their own internal ethical guidelines and procedures that researchers must follow". Ten years on, this hypothesis seems to hold, particularly given that most explicit discussions of ethical considerations in corpus linguistics seem to occur in contexts where corpora representing communicative contexts in which standards for ethical practice are apparently less well established and still debated. For example, the particular ethical considerations underpinning specific corpus linguistic research projects seem to be receiving greater attention in research publications reporting on studies of online, social media interaction (e.g. Brookes 2018; Wright 2020; Jones et al. 2022). Meanwhile, Collins (2019) and Hunt and Brookes (2021) provide detailed discussions of ethical considerations which are likely to apply to corpus studies of computer-mediated communication in a more general sense. These exceptions aside, there remain few examples of explicit discussion of ethical considerations underpinning corpus linguistics research and practice (for further discussion of online research ethics, see Chapter 5).

This chapter responds to the need for more detailed discussion of such issues by providing a critical exploration of ethical considerations in corpus linguistics. We build on what is, in our view, the most detailed exploration of ethical issues in corpus linguistics to-date (McEnery & Hardie 2012). Specifically, like McEnery and Hardie (2012), we will approach all

ethical issues under discussion from the perspectives of those involved in corpus linguistics research and practice. Accordingly, we organise our discussion according to the following perspectives: i.) research participants; ii.) corpus builders; iii.) corpus distributors; and iv.) corpus users. Of course, the aforementioned perspectives are not discrete (i.e. a single person will frequently perform several of these roles within a single project), and while certain ethical issues are particularly pertinent to one role or another, in reality they cut across and have implications for various aspects of corpus design and use. However, structuring our discussion in this way is helpful both for organising our thinking and for guiding ethical considerations, broadly as these arise through the typical chronology of a corpus linguistic project. Most of the issues we discuss over the coming pages are not specific to ethical considerations in corpus linguistics but interleaf with those in Applied Linguistics generally, as well as other sub-branches of the field specifically. That said, we will devote most of our space to discussing ethical issues that are, in our experience and based on our knowledge of others' work in the field, particularly pronounced within corpus linguistics.

Ethical Considerations in Corpus Linguistics

Research Participants

While research ethics can be viewed from a range of perspectives (see De Costa, 2016), the 'protection of human subjects, including obtaining participant consent … and maintaining confidentiality of research data, is perhaps the prototypical concern in research ethics and certainly an important one' (Isbell et al. 2022, p. 173; see also Sterling & Gass 2017). For the purposes of this chapter, we regard participants as those people who originally produced the

language that is included in a corpus. Broadly, professional organisations and institutional review boards typically advise that informed consent should be obtained from participants in any study before their language is collected and their contributions included in the corpus. This now represents standard practice in Applied Linguistics research. Unlike in areas of Applied Linguistics where language data is typically elicited in laboratory or experimental conditions, corpus linguistics' commitment to analysing naturally occurring language use means that corpus compilers engage with participants from a wide range of real world contexts of language use, and this can make the process of contacting participants and obtaining informed consent more difficult compared to, say, working with participants in a laboratory experiment.

To demonstrate this point, we can consider the influence that the mode and medium of the language collected for a corpus can have over ethical considerations. For the collection of written texts, it can be important to obtain the informed consent of those who composed the texts in question, particularly if those texts were not intended for public consumption, such as private letters or essays written by students. For written texts produced for public consumption, such as novels, newspapers, and academic journal articles, informed consent is not typically required, though copyright restrictions may mean that permission has to be sought from the copyright holder before a text can be included in a corpus. For the compilation of corpora of written texts, it is standard practice in the field to obtain informed consent from participants *after* the text has already been written, but *before* it is included in the corpus, unless there is legal provision that does not require that (Brookes & McEnery 2021). For example, Brezina et al (2021) discuss how the provisions of copyright legislation in the UK permit the use of electronic texts for research as of right, with professionally published texts being produced in full knowledge of these legal provisions, making the seeking of further permission un-necessary.

For the compilation of spoken language corpora which contain recordings of speech, informed consent will have to be obtained from the participants prior to the recording taking place, as happened, for example, with the production of the Spoken BNC 2014 (see Love et al., 2017). As with the collection of written language, the distinction between public and private domains is an important consideration; spoken language produced for a public audience will typically not require informed consent to be obtained and analysed. However, again, copyright restrictions may have to be taken into consideration for the collection of such texts.

Perhaps the most challenging mode of language, ethically, for corpus linguists to collect is online, computer-mediated communication. Any practical benefits of having such born-digital language readily available in an electronic, corpus-friendly format are often swiftly counterbalanced by the lack of standardised ethical practice regarding such data. Ethical guidance regarding online research increasingly acknowledges the variability in the forms and contexts of online communication and, as such, reject 'blanket policy' approaches in favour of process-driven ones. According to such process-driven approaches, ethical decisions are made on a case-by-case basis and should respond to each stage of a study (identifying a potential research field, accessing the site, gathering data, publishing results, etc.; Hunt & Brookes 2020).

One of the biggest challenges facing those designing and assembling corpora of online language concerns the unclear distinction between private and public domains. The latter is widely accepted to be more likely to require informed consent from participants. Yet discerning between private and public domains is not always straightforward when dealing with online communication, as Hunt and Brookes (2020, p. 72) discuss:

> From a technical perspective, many online contexts – and particularly websites whose content is predominantly user-generated – are freely accessible to the public, with their

> content comparable to a publicly authored digital book […]. From this perspective, web users have opted to communicate in a public setting and their data can be used for research purposes without their notification or consent. However, even while technically public, web users may nevertheless perceive their interactions as at least partially private and can suffer harm when information that they perceived to be private is published in different contexts (Frankel & Siang, 1999).

To address this, some researchers have employed Nissenbaum's (2010) model of 'contextual integrity' as a way of understanding the norms and expectations held by members of online communities regarding the use and flows of their information (see, for example, Mackenzie's [2017] ethnographic study of Mumsnet). However, doubts have been raised as to the feasibility of corpus compilers and analysts achieving contextual integrity, where the size and scope of their data is likely to represent a larger number of unknown (and unknowable) participants than is typically the case for ethnographic studies (see Collins 2019). This challenge becomes more pronounced when researching communities of which the corpus compilers are not themselves a part. In another approach, Giaxoglou (2017), following Page et al. (2014), offers a matrix for determining the level of risk that a study's design poses to participants (see Table 1).

Table 1: Dimensions of research and level of risk (adapted from Giaxoglou, 2017).

Dimensions of research	Low-risk	High-risk
1. Types of data	Large scale or 'big' data obtained via computerized	'Small' data obtained via ethnographic observation

	programmes (e.g. java, API protocols)	methods, interviews, surveys, online ethnographies
2. Methodology	Quantitative	Qualitative
3. Site/platform	Sites/platforms with privacy settings (e.g. Facebook)	Sites/platforms without privacy settings (e.g. early days of MySpace)
4. Research focus	Focus on large-scale trends Focus on discourse patterns Focus on texts	Focus on persons and their lives

A study situated closer to the higher risk characteristics along these dimensions of research should, it is suggested, necessitate greater researcher precaution. However, again, corpus studies might not fit neatly onto these dimensions, as Hunt and Brookes (2020) point out, since corpus analyses often involve both quantitative and qualitative elements, and corpora may include multiple sites of online interaction with varying privacy settings.

Existing frameworks guiding linguistic research are thus more ideally suited to qualitative (often ethnographic) studies and may be limited for addressing the complexities involved in online corpus compilation. For example, Hunt and Brookes (2020) gathered data from websites which could be deemed to be public as they did not require registration or permission to access. It was also stipulated that the sites should contain visually prominent user

guidelines which emphasized the public nature of the forum and/or the breadth of people that may read users' contributions and did not prohibit research from taking place on the fora. They also stipulated that the fora should have a large number of members (ranging, in their case, between several hundreds and tens of thousands), which is argued to correlate with contributors' perceptions that their messages will have a wide-ranging audience (see Eysenbach and Till 2001). Another feature of the fora taken as a signal of their public nature was the availability of other means of communication within the sites for users to communicate more privately (e.g. direct messages and password-protected sub-fora). Following Giaxoglou (2017), who recommends analysing online contributors' discourse to understand their orientation to the context as more private or public, Hunt and Brookes (2020) observed how the contributors in their study typically oriented to a wider audience, as indicated through the adoption of plural first-person pronouns to speak on behalf of the forum, as well as addressing thread-initial posts to 'anyone' or 'everyone'. On these bases, Hunt and Brookes (2020) judged the sites they were studying to be more public than private.

The corpus compiler's commitments to their participants extends beyond considerations around informed consent, though, and another consideration which cuts across all corpora of all modes of communication relates to participant anonymity. Generally speaking, approaches to corpus compilation emphasise the importance of anonymising the language included in a corpus to protect the identities, of the participants involved. As a consequence, corpora of written, spoken and digital interactions are typically subjected to a process of de-identification, whereby references to names and other personal details by which participants and third parties might be identified are manually modified or removed. This process is particularly important if the corpus is to be made publicly available, as was the case for the aforementioned BNC 2014, but for

corpora which are not to be shared beyond the immediate research team it can be sufficient to anonymise data extracts reproduced for public consumption (e.g. in research outputs) rather than having to anonymise the corpus itself (though the increasing movement towards 'open science' and data sharing practices in corpus linguistics is likely to make the former approach more commonplace in the future; see Chapter 1).

As well as the purposes to which the corpus will be put, another factor influencing considerations around participant anonymity concerns the mode and context of the language represented in the corpus. For written language, if the texts sampled are freely accessible in the public domain, then there is not usually an expectation that they will be anonymised. If the texts are not publicly available, then the texts may need to be anonymised, in line with what was agreed with participants or copyright holders as part of the informed consent process. Modifying or removing identifying information from written texts can be labour-intensive but is not particularly complicated. This is also the case for constructing corpora of spoken transcripts or recorded speech.

Preserving anonymity is much more complex when dealing with audio or video recordings of conversations. This is, as Knight and Adolphs (2021, p. 27) point out, because:

> Audio data is "raw", existing as an "audio fingerprint" of vocalisations that are specific to an individual…. A similar problem arises with the use of video data. Although it is possible to shadow, blur or pixelate video data in order to conceal the identity of speakers, these measures can be difficult to apply in practice (especially with large datasets).

Knight and Adolphs (2021) suggest that it is important to discuss such anonymisation issues with participants prior to recording, to ensure that they understand what the recordings will capture, as well as how these will be used and distributed.

Regarding the compilation of corpora of online language, even if we judge a particular online communicative context to constitute a public rather than private domain, corpus compilers should be mindful that participants may not have envisaged their language being shared as part of a widely distributed corpus or reproduced as data extracts within academic publications. Here, the searchability of online texts presents an additional challenge, and one which may become even more pronounced when sampling language that is produced around sensitive topics or by groups we might perceive as vulnerable (discussed more below). For this reason, in the aforementioned study of online interactions around mental health and distress, Hunt and Brookes (2020) used pseudonyms to refer to the fora from which they sampled threads of messages and removed all identifying information (including contributors' online usernames) from quoted data. They also collated corpora from fora which, at the point of data collection, were not indexed on search engines. This measure helped to ensure that the participants' online profiles could not be identified using quoted data. In addition to the technical aspects of the platforms from which online language is gathered, we should also take into account the nature of the communities themselves. This is something we discuss in more detail in the next section.

Corpus Builders

While many of the issues discussed in the previous section have implications for corpus builders and the decisions they make, some ethical issues impact corpus builders more directly. An example of this is provided by McEnery and Hardie (2012), who describe how, when constructing a parallel corpus of English aligned with a number of South Asian languages as part

of the construction of the EMILLE corpus (Baker et al. 2004), they were working in an opportunistic mode. This was because there was little available data covering all of the languages they had intended to include in the corpus. They were approached by a religious organisation which wanted to help; this organisation had translated many of their English-language magazines and leaflets into the languages that the research team were struggling to obtain for the corpus. Although this represented an attractive opportunity to expand the corpus, the researchers felt they had to decline their offer. This was partly because the organisation in question saw the corpus as a way of distributing their material and thus gaining converts. However, the researchers were uncomfortable with the idea of their research work becoming missionary work. Furthermore, when they surveyed the material itself they found it to be offensive; for example, one magazine ran an article entitled 'Who it is alright to hate' (sic). Since the aim of the research team was not to construct a corpus of missionary texts, nor to construct a corpus of morally censorious texts, they decided that the texts in question compromised their ethical stance and thus rejected the offer of the data.

McEnery and Hardie (2012) also point out that while the solution to their ethical dilemma was clear in this case, the situation may have been complicated had religious texts been part of their corpus sampling frame (as is the case with the Brown sampling frame, for example). In that case, rejecting material on grounds of offensiveness may have resulted in skewing the texts in the corpus towards particular philosophical or theological perspectives. In this case, the research team did not feel that this skew was an issue, as they were aiming only to compile samples of official documents (broadly defined). Corpus builders may find themselves in the same position as Baker et al. (2004), where they are forced to confront an issue such as this. That is because, as McEnery and Hardie (2012, p. 65) point out:

> the underlying problem is embedded in one of the great strengths of the corpus approach: corpora are multifunctional. Once built, they may be used for a wide range of purposes, some of which the builders of the corpus would never have imagined, and quite possibly some which they would never have approved of.

Sometimes, as well as the ideology of the material sampled for a corpus, corpus builders may have concerns around researching or accessing particular communities, particularly those which cohere around ideologies or practices that might be perceived as violent, discriminatory or which are otherwise considered unpalatable to the researcher. For example, Rüdiger and Dayter (2017) explored the ethical conundrums involved in carrying out linguistic research on online platforms populated by 'pick-up artists' – a community that 'learns and practices speed-seduction for short-term mating' (2017, p. 251; see also Wright 2020). These researchers explore the influence that researching such hostile communities might have on traditional methodological decisions. This case study underscores the importance of considering the nature and functions of the communities under study when making decisions around ethics – in this case, we should consider the possibility that researchers may place themselves at personal risk by reaching out to likely hostile groups which cohere around discriminatory practices.

Another way in which ethical issues may attach directly to corpus builders concerns the (perceived) impact that the corpus builder may have on members of the communities they are accessing for the purposes of data gathering, particularly where this involves obtaining informed consent. Researching online discussions of sensitive topics is likely to render more acute the need for researchers to be responsible, respectful, and sensitive in their approach to how they collect, use, and report on the language produced by such groups. While such considerations are, of course, consistent with the general principles guiding ethical research practice, they can also

introduce additional complexities for the corpus compiler. Elgesem (2015) urges those researching online communities that cohere around the discussion of sensitive health-related topics, to weigh the need for informed consent against the implications of obtaining it. With this in mind, Hunt and Brookes (2020) judged that requesting informed consent from a large number of mental health forum users might risk disrupting the fora and undermining their primary role as recovery-oriented communities (see also Nosek et al. 2002). They judged this risk to be greater than the potential harm that would likely be brought about by the passive observation of the fora they undertook, though they took additional steps to preserve the anonymity of the forum participants who featured in their corpora.

The case studies discussed above highlight the complexities involved in judging the vulnerability of research subjects, in evaluating the potential harm in linguistic research, as well as the limitations of employing blanket policies in relation to informed consent practices (particularly where such online communities might also be considered 'public').

Corpus Distributors

Building on the ethical considerations surrounding corpus construction are those surrounding its distribution. With the move towards open science and data sharing, corpus linguists are increasingly considering how to create corpora in ways that are not only ethically sound, but also usable by other researchers. Indeed, since study replicability is a key goal of corpus linguistics, corpus distributors have a responsibility to ensure that their data remain intact and available for other analysts in the future (where ethically and commercially permissible). Yet in addition to this, in the spirit of 'open science', we all have an ethical obligation to share not only corpora

(when ethically appropriate) but also any code, scripts, and other protocols we use in the construction and analysis of corpora. If a corpus is to be made publicly available, or even distributed more widely than the immediate research team, then participants will have to be made aware of plans for distribution before they can give informed consent. A further ethical challenge associated with such widely distributed corpora is that it can be difficult or even impossible to deal with participants' requests to retract their consent (and to remove their contributions) once the corpus has been shared.

Yet beyond the ethical imperative to make data available (where possible), corpus distributors may face other ethical challenges. Unless access to the data is available on a request-only basis, corpus distributors cannot realistically control what others will do with their corpora. However, they could make it clear in a corpus manual what the ethical conditions of the corpus collection were, including its original intended purpose. Any publicly available corpus may be used, quite legally and legitimately, by organisations which carry out or commission research whose aims are to meet to objectives of that organisation. If we, as corpus distributors, are sympathetic or ambivalent towards those organisations and their motivations, this may not pose any ethical challenge to us. However, we may find those organisations and their motivations more objectionable; an example of this is the offer of religious texts to Baker et al. (2004), where accepting the offer of those texts would have entailed a religious use of their corpus distribution that the research team did not intend nor desire. Such cases demonstrate the kinds of ethical issues that can be confronted by corpus distributors, and which they might want to bear in mind when establishing the rules for redistribution and future use. We would advise corpus users to contact corpus distributors if they are unsure of the ethical grounds of their reuse of corpus data.

Corpus Users

Finally, corpus users are also likely to encounter a number of ethical considerations in how they access and analyse corpora, as well as how they report their findings. Pimple (2002) outlines a series of components of ethical research practice, three of which are identified as being particularly relevant to Applied Linguistics by Sterling and De Costa (2018). These principles can be broadly mapped onto some of the main considerations underpinning ethical corpus use. First is the wisdom to only conduct research that is meaningful and useful to society, and which does not involve researchers brining about unnecessary harm. For corpus linguists, this can pose a challenge, as we may be tempted by the ready availability of a large amount of language data that is almost 'corpus-ready'. Compiling and analysing such data might not be problematic, but if such a project involves gathering language produced by individuals – particularly if those individuals might be considered vulnerable, their language concerns some sensitive topic, or the context of language use is ethically complex – then we should have a clearly defined purpose for that corpus which will bring some benefit to society, other researchers and/or, ideally, the language users under study themselves. For applied corpus linguistics research, this can mean carrying out research which has pedagogical implications, but it might also involve furthering our knowledge about language use in a more general sense (Sterling and De Costa 2018).

The second of Pimple's (2002) components of ethical research that is relevant here is a commitment to truth in the reporting and representation of data. Here, the principle of total accountability, originally advocated in corpus linguistics research by Leech (1992), is relevant. That requires, that a) all available data is gathered for analysis (i.e. texts are not excluded on the grounds that they contradict a pre-existing argument or theory), and b) all data gathered must be accounted for. While corpus linguistic techniques provide analysts with the tools to undertake

arguably more objective analyses than is likely possible using purely qualitative methods of linguistic analysis, it is important to acknowledge that corpus linguistic methods are not 'objective'; as this chapter itself has attested, humans are required to make important decisions at each step in a corpus linguistic study – from design, to data collection, to analysis and finally reporting of findings. Being transparent at each step, in terms of describing what decisions were made and why, can help researchers to be more accountable. When it comes to the analysis, corpus output cannot be interpreted by a computer – this is a task which must be undertaken by the human analyst. Corpus analysis can be challenging, as we are typically dealing with large volumes of data and the procedures we carry out can yield lots of results – more than can feasibly be reported within the scope of a journal article and, in many cases, even an entire monograph. We therefore have to exercise some selectivity about the patterns and results we report. Here, corpus users should resist any temptation to 'cherry pick' results that are interesting, or which confirm our hypothesis, particularly at the expense of reporting those which offer counterexamples. Again, making our data available to others can help in this process, as can actively searching and reporting on the frequency of counterexamples (i.e. patterns which run counter to the dominant trend or which do not fit our hypotheses (see Partington et al. 2013). Our interpretation of the corpus output is also likely to be influenced by our subjective experiences and potential biases – being reflexive about this and acknowledging it when reporting our results represents good ethical practice for corpus users.

Describing what we, as analysts, do to obtain results, and acknowledging how our subjectivities may have influenced these, is clearly good scientific and ethical practice. Yet this practice can be complicated further when using corpus linguistic methods – and other quantitative techniques – as our analyses may be based on algorithms embedded within the

software we use. Maintaining such software and providing clear and detailed records of the procedures it operates by are important for ensuring transparency and replicability. If software is updated regularly, this means that a particular analytical procedure could be carried out on the same dataset but yield different results over time, or indeed at a single point of time, depending on the version of the software being used. It is therefore the duty of developers to provide detailed and up-to-date records of modifications to the software they release, and the duty of corpus users to report which version they are using. The same principle applies to the use of corpora themselves, which can be released in different iterations and, in some cases, updated over time.

Beyond transparency in reporting our methods and results, it is also incumbent on corpus analysts to ensure the scientific validity of their research. This essentially ensures that the research itself is carried out in a methodologically rigorous way, and that any claims to statistical salience are robust and accurate. Methodological rigour may not seem, at face value, to be an ethical issue. However, results from linguistics research can inform the policies or actions of powerful organisations, such as governments, which can in turn affect the lives of millions of people (De Costa, 2018). In a similar vein, it is also the ethical responsibility of corpus users to reflect critically on what their results represent, and to exercise similar criticality when suggesting what the implications of their findings are and issuing recommendations based on these. A relevant example relates to work examining patient feedback on UK healthcare services (Baker et al., 2019). One of the findings was that receptionists were more likely than other types of staff member to receive negative evaluations in the patients' feedback (Brookes & McEnery, 2020; Baker & Brookes, 2021). When reporting that finding to stakeholder contacts in the healthcare system, the researchers felt that it was important to contextualise that pattern, and to

foreground the fact that the feedback was shaped by patients' expectations, as well as the unique role of receptionists (being gatekeepers and the public faces of systems they don't design). In other words, it was not the case that receptionists were poor at their jobs, but that patients were often not sufficiently aware of what receptionists' roles entailed, and that their criticism of receptionists was often, in fact, personalised criticism of broader organisational issues within the healthcare system. We have to be responsible with our findings, then, and mindful of how they might be interpreted, and misinterpreted, and guard against any undue harm arising from such misinterpretation. Engaging with perspectives on the given issue from other disciplines, and indeed engaging with members of the group concerned directly, can help us to make sense of our findings and bring better balance and fairness to our conclusions.

Another feature of interpretation relates to what we claim based on our measurements in a corpus – supporting claims with statistics which are appropriate for the claim in question is clearly important. Perhaps the most salient example of this is the use of effect size measures. While significance statistics may have a role to play in asserting a difference between two corpora, for example, they remain largely silent on the question of how great that difference is. Measuring the scale of difference is where effect size measures are important and should be used where claims of differences in scale are made. See Brezina (2018, p. 20) for a discussion of how the different measures available to us articulate together.

Pimple also urges that ethical research is fair when citing and using the work of others. Fairness is, of course, a good principle for researchers to follow in general, regardless of what field or sub-field they are working in. For corpus users, the issue of citing resources (i.e. corpora, analytical tools, etc.) is of particular relevance, since most of us in the field rely on the innovations and work of others in developing and maintaining publicly available corpora and

computer packages which improve the efficiency and accuracy of our analyses. It is important that we acknowledge the creators of such resources whenever we use them. Corpus project are often group projects, and they increasingly involve researchers from other fields too. It is ethical practice to acknowledge the work of all involved in any project, and to be clear about aims and expectations from all parties from the outset of collaborative projects.

Conclusions

In this chapter, we have addressed ethical issues, giving particular focus to those issues that are likely to be of pronounced importance to researchers embarking on corpus linguistics projects. The ethical considerations underpinning any corpus study are influenced by the kind of language under study, the real-world contexts in which it occurred, the methods and type of analysis being employed and, perhaps most importantly, the researcher's subjectivity and own ethical positions. For this reason, the kind of critical discussion of ethical considerations we have engaged in here is, we hope, likely to be of greater use than any blanket guidelines which will not realistically be able to reflect the wide and ever-growing range of modes, contexts, and groups that are represented in corpora. Throughout this chapter we have issued a series of specific, though rather general, recommendations for addressing some of the specific ethical issues that we and others have encountered in previous research. In this section, we will briefly reflect on where we are, now, as a field, and consider how future engagement with ethical considerations in corpus linguistics might proceed and add value to the field.

Ethically problematic research design in corpus linguistics is, we feel, largely a thing of the past and, even then, such practices likely reflected the infancy of the field relative to other

areas of Applied Linguistics. Lessons have certainly been learnt since then, and such practices are thankfully confined to the past. For example, as a case in point, we can contrast the covert approach to recording taken in the compilation of the original British National Corpus, which involved secretly taping the speech of unknowing corpus contributors, against the ethically robust 2014 update, which required informed consent to be obtained from contributors *prior to* their speech being recorded for inclusion in the data. Nowadays, the kinds of ethical practices that are becoming standard in corpus linguistics research are, we feel, compatible with other areas of Applied Linguistics. Corpus builders take the privacy concerns of participants seriously, even when confronted by a lack of guidelines or 'gold standard', and do not routinely produce overtly unethical corpora or study designs. Corpus distributors go to great lengths to ensure the data they distribute confirms with every-changing legal and ethical standards, and invest much time and effort into maintaining those standards and producing documentation to enhance the accessibility and transparency of their resources. Now more than ever before, corpus users explore sensitive issues – for example, relating to health, crime, and extremism – in a sensitive, nuanced, and ethically responsible way.

The above point notwithstanding, there has been very limited explicit engagement with ethical issues in corpus linguistics specifically. Such engagement with ethical considerations in corpus linguistic is therefore needed, particularly as the volumes and types of data available to corpus linguists will continue to expand at the pace. We need to be reflective and open to interdisciplinary research meaning that ethics norms or standards will need to be negotiated across disciplinary boundaries. The need for that will become more pronounced as corpus linguistics reaches into ever more (and seemingly more distant) disciplines beyond linguistics, for example within the social sciences (see: McEnery & Brookes, fc.).

While we observed the general paucity of discussion of ethics in major introductory textbooks and handbooks in corpus linguistics, what is likely to be of greater value to garnering a greater appreciation of the situatedness of ethical considerations is explicit, detailed, and honest reflection on the ethical issues considered in individual corpus studies. Ethically reflective contributions, such as those by Rüdiger and Dayter (2017) and Mackenzie (2017), demonstrate the value of sharing lessons learnt from specific projects for the benefit of others. Similar endeavours focusing on corpus linguistics research specifically could help us to better understand contemporary ethical practice in the field, and initiate productive debate and exchange of practice.

In terms of research practice itself, an important ethical consideration for corpus linguists to engage with, particularly those working on topics of particular social relevance, is the extent to which the common practice of focusing on frequent and statistically salient forms of discourse and language use might amplify the practices and perspectives of dominant groups while serving to further background those of the less powerful. This is a social justice issue that Applied Linguistics is confronting in general. It is also one that corpus builders and analysts should reflect on in terms of what (or rather, who) their corpora represent, and what might be lost when we our analyses focus on frequent (and, thus, typically dominant) voices in society.

References

Baker, P., Hardie, A., McEnery, A., Xiao, R., Bontcheva, K., Cunningham, H., Gaizauskas, R., Hamza, O., Maynard, D., Tablan, V., Ursu, C., Jayaram, B. D., & Leisher, M. (2004). Corpus

linguistics and South Asian languages: Corpus creation and tool development. *Literary and Linguistic Computing*, *19*(4), 509-524.

Baker, P., & Brookes, G. (2021). Lovely nurses, rude receptionists, and patronising doctors: Determining the impact of gender stereotyping on patient feedback. In J. Angouri & J. Baxter (Eds.), *The Routledge handbook of language, gender and sexuality* (pp. 559-571). Routledge.

Baker, P., Brookes, G., & Evans, C. (2019). *The language of patient feedback: A corpus linguistic study of online health communication*. Routledge.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.

Biber, D., & Reppen, R. (2015). *The Cambridge handbook of English corpus linguistics*. Cambridge University Press.

Brezina, V. (2018). *Statistics in corpus linguistics*. Cambridge University Press.

Brezina, V., Hawtin, A., & McEnery, T. (2021). The Written British National Corpus 2014 – design and comparability. *Text and Talk*, *41*(5-6), 595-615.

Brookes, G. (2018). Insulin restriction, medicalisation and the Internet: A corpus-assisted study of diabulimia discourse in online support groups. *Communication & Medicine*, *15*(1), 14-27.

Brookes, G., & McEnery, T. (2020). Corpus linguistics. In S. Adolphs & D. Knight (Eds.), *The Routledge handbook of English language and digital humanities* (pp. 378-404). Routledge.

Brookes, G., & McEnery, T. (2021). Building a written corpus: What are the basics? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 35-47). Routledge.

Collins, L. (2019). *Corpus linguistics for computer-mediated communication*. Routledge.

De Costa, P. (2016). *Ethics in applied linguistics*. Routledge.

Elgesem, D. (2015). Consent and information – Ethical considerations when conducting research on social media. In H. Fossheim & H. Ingierd (Eds.), *Internet research ethics* (pp. 15-34). Nordic Open Access Scholarly Publishing.

Eysenbach, G., & Till, J. E. (2001). Ethical issues in qualitative research on internet communities. *British Medical Journal*, *323*, 1103-1105.

Frankel, M. S., & Siang, S. (1999). *Ethical and legal aspects of human subjects research on the internet: A report of an AAAS workshop*. American Association for the Advancement of Science.

Giaxoglou, K. (2017). Reflections on internet research ethics from language-focused research on web-based mourning: Revisiting the private/public distinction as a language ideology of differentiation. *Applied Linguistics Review*, *8*(2-3), 229-250.

Hunston, S. (2022). *Corpora in applied linguistics* (2nd ed.). Cambridge University Press.

Hunt, D., & Brookes, G. (2020). *Corpus, discourse and mental health*. Bloomsbury.

Isbell, D. R., Brown, D., Chen, M., Derrick, D. J., Ghanem, R., Arvizu, M. N. G., Schnur, E., Zhang, M., & Plonsky, L. (2022). Misconduct and questionable research practices: The ethics of quantitative data handling and reporting in applied linguistics. *The Modern Language Journal*, *106*, 172-195.

Jones, L., Chałupnik, M., Mackenzie, J., & Mullany, L. (2022). 'STFU and start listening to how scared we are': Resisting misogyny on Twitter via #NotAllMen. *Discourse, Context & Media*. Online first.

Knight, D., & Adolphs, S. (2021). Building a spoken corpus: What are the basics? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 21-34). Routledge.

Leech, G. (1992). Corpora and theories of linguistic performance. In J. Svartvik (Ed.), *Directions in corpus linguistics* (pp. 105-122). Mouton de Gruyter.

Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). The spoken BNC 2014. *International Journal of Corpus Linguistics*, *22*(3), 319-344.

Mackenzie, J. (2017). Identifying informational norms in Mumsnet Talk: A reflexive-linguistic approach to internet research ethics. *Applied Linguistics Review*, *8*(2-3), 293-314.

McEnery, T., & Brookes, G. (Forthcoming). Corpus linguistics and the social sciences. *Corpus Linguistics and Linguistic Theory.* In Press.

McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.

McEnery, T., & Wilson, A. (2001). *Corpus linguistics: An introduction* (2nd ed.). Edinburgh University Press.

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). E-Research: Ethics, security, design, and control in psychological research on the internet. *Journal of Social Issues*, *58*(1), 161-176.

Page, R., Barton, D., Unger, J., & Zappavigna, M. (2014). *Researching language and social media: A student guide*. Routledge.

Partington, A., Duguid, A., & Taylor, C. (2013). *Patterns and meanings in discourse: Theory and practice in corpus-assisted discourse studies (CADS)*. John Benjamins.

Pimple, K. D. (2002). Six domains of research ethics: A heuristic framework for the responsible conduct of research. *Science and Engineering Ethics*, *8*, 191-205.

Rüdiger, S., & Dayter, D. (2016). The ethics of researching unlikeable subjects: Language in an online community. *Applied Linguistics Review*, *8*(2-3), 251-269.

Sterling, S., & De Costa, P. (2018). Ethical applied linguistics research. In A. Phakiti, P. De Costa, L. Plonsky, & S. Starfield (Eds.), *The Palgrave handbook of applied linguistics research methodology* (pp. 163-182). Palgrave Macmillan.

Sterling, S., & Gass, S. (2017). Exploring the boundaries of research ethics: Perceptions of ethics and ethical behaviors in applied linguistics research. *System*, *50*, 50-62.

Wright, D. (2020). The discursive construction of resistance to sex in an online community. *Discourse, Context & Media*, *36*, 1-9.