3 The web, laws and ethics

3.1 Introduction

In this chapter we will continue our survey of practical issues that may arise when working with corpus data. While the first two chapters discussed the selection, annotation and exploitation of the data in a corpus, in this chapter we will consider two important and related issues: legal considerations in corpus construction; and the equally important, yet less often discussed, ethical issues arising from corpus construction, distribution and use. We will begin with the legal issues, which have become more pressing over time as vast amounts of textual data have become available to collect easily over the World Wide Web. Accordingly, in this chapter we will approach legal issues in corpus construction with specific reference to compiling corpora from the web. In doing so, we will also consider some of the practical issues around web-based corpus construction.

We should note that we write, in this chapter, very much from the standpoint of Western culture. Laws and ethics vary across the planet – but rather than attempt a global survey, we seek here to illustrate the relevant legal and ethical issues from the context in which our own research is undertaken.

3.2 The web and legal issues

The most fundamental issue in corpus construction is whether or not you have the legal right to gather and distribute the data you intend to include in your corpus.

The massive expansion of the World Wide Web in the mid-to-late 1990s presented both opportunities and problems for corpus builders. Before the age of the web, to collect a text in electronic form it was necessary either to get the original file from the publisher, or to rely on re-typing (time-consuming and expensive) or optical character recognition software (error prone). However, the hypertext documents that make up the web are already in electronic form, and frequently in an encoding and format (ASCII text with HTML markup) very similar to the XML format preferred for corpus data. Thus, it has become...
the web, laws and ethics

extremely straightforward simply to download and save large quantities of text from the web to create a corpus – either manually, or for a larger corpus using an automated program called a web crawler. One such automated program which is specifically designed for linguistics is BootCat (Baroni and Bernardini 2004). This program is ‘a suite of Perl programs implementing an iterative procedure to bootstrap specialised corpora and terms from the web, requiring only a small list of “seeds” (terms that are expected to be typical of the domain of interest) as input’ (Baroni and Bernardini 2004: 1313). Given the availability of such tools, it is hardly surprising that the study of the ‘Web as Corpus’ (which we have introduced already in section 1.4.2) has become a highly active subdiscipline of corpus and computational linguistics, with some studies focusing upon genres unique to the web, e.g. online chat rooms (Claridge 2007; Thelwall 2008; King 2009). However, while there are linguists who point out the obvious attractions of the web, conceiving of it as ‘a fabulous linguist’s playground’ (Kilgarriff and Grefenstette 2003: 333), there are others who urge caution, noting that the web ‘can in no way be considered a representative sample of language use in general’ (Leech 2007: 145). It is unsurprising then that some have concluded that 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).

While being able to use the web is a great advantage for corpus construction, there are also problems. One issue is the difficulty of determining the genre of any given document harvested from the web without actually reading it – knowing the genre is necessary for a balanced corpus design. Though BootCat tries to do this, its success is heavily dependent on the user being able to select terms for their search which are strongly associated with the genre in question. An initial evaluation by Baroni and Bernardini (2004: 1315) suggests that one in three of the webpages recovered may not be in the desired genre. However, while the usefulness of the web for linguistics may be debatable, and while the programs, such as BootCat, can be improved, one constant remains – the legal issues are as complex for large-scale web harvesting as for any other type of corpus construction. Copyright laws apply to documents available on the web exactly as they do to print documents. That is, it would be illegal to download a text from the web and then redistribute it as part of a corpus without the permission of the author of the webpage. While this may seem unreasonable, given that the majority of websites are entirely open to public view, many content providers on the web are reliant on fees for advertising that are paid per visitor. So if even one person who might have looked at the original webpage instead sees a copy in a corpus, the creator of the original content suffers a financial loss. There can be no objection to someone downloading a single copy of a document on the web onto one computer for their own use (indeed, such copying happens every time a web browser visits a page). But it would clearly breach copyright to redistribute these local copies. This is a serious problem: just as corpus tools need to be widely available if corpus linguistics is to be replicable, the actual corpus data also needs to be made as widely available as possible, for
3.2 The web and legal issues

precisely the same reason. This is, we would argue, an ethical imperative for the researcher.

There are several ways of addressing the copyright issues around data collected from the web. The first is to treat text from the web in the same way as any other text. That is, the corpus builder contacts the copyright holder and requests permission to redistribute the text within a corpus under the terms of some specified licence. This procedure was followed by the builders of pre-web corpora such as the BNC and LOB, but also by the builders of the EMILLE corpora, for which much of the data was gathered from selected websites (see Baker et al. 2004). This is feasible if one or a small number of websites are to be sampled. This need not mean a small corpus – many news sites, blogs and internet forums contain many millions of words of archived data. But it would certainly lead to a corpus that represents only a narrow slice of the variety of the web. While this might be a perfect dataset for some sets of linguistic questions, it would be of rather less use for addressing questions about a given language in general. An alternative is to collect data only from sites which explicitly allow the reuse and redistribution of text. A website may declare that its content is public domain, or that it is available under a licence which permits copying and redistributing. This is increasingly common. For example, all pages on the multilingual Wikipedia site are licensed as ‘free documentation’, meaning that making and distributing copies is permitted (with some conditions). However, restricting a corpus to such sites would again skew its representativeness.

The third approach is to collect data without any regard to seeking permission, and not to distribute it, but instead to make it available to other researchers through a tool that does not allow copyright to be breached. Many web corpora are made available through fourth-generation, web-based concordancers (see section 2.5.4) where only a few words of context around the node word are visible. Indeed most web search engines might be viewed as primitive concordancers of this sort. Since it is impossible to reconstruct the original texts from the tiny snippets in the concordance, which are small enough to count as ‘fair use’, this ‘redistribution’ does not constitute a dangerous copyright violation. Thus, the corpus is available to other researchers, since they can run the same searches from the same web-based tool, but the law has not been broken. This is less than entirely ideal, however, because some more sophisticated corpus analyses (especially those involving advanced statistical calculations, e.g. collocation strength across variably defined subcorpora) cannot be done without access to the text of the entire corpus. Davies, in a discussion of such online interfaces, acknowledges this problem, but points out that ‘in many cases it is still possible to obtain large amounts of word frequency and n-gram information from these corpora’ (Davies 2010: 417) as well as perform advanced lexical and grammatical searches. As fourth-generation concordancers such as Davies’ interface continue to be improved and developed, the analyses they permit will be continually extended. But, inevitably, a web-based concordancer will never allow the full range of analyses that a technically savvy researcher could accomplish with a copy of
the corpus on their own computer. Moreover, fair-use law – like all parts of copyright law – may vary between countries (and across legal jurisdictions within countries). So far as we are aware, there has been no definitive establishment of the legal status of fair-use-snippet web concordancers – for instance, no test case has been taken to court that we know of. There are several reasons for this: those who create web concordancers are very careful on legal issues; corpus linguistics is fairly obscure in the grand scheme of things and most text producers probably don’t even know if their text ends up in a corpus that is searchable online; and, ultimately, corpus linguists are unlikely to have enough money to be worth suing.

A more innovative solution to the copyright issue is to redistribute not the downloaded data files, but rather a list of the web addresses from which the corpus has been collected (and, if necessary, details of the procedures used to download and process the texts). This does not breach copyright at all – but any researcher with the appropriate software can download those webpages and reconstruct their own personal copy of the corpus from the address list quickly and easily. This ensures that any findings from such a corpus are open to being checked and replicated, at least as long as those pages remain available and unchanged on the web. However, the permanence of webpages is, of course, highly variable; so while this approach has some advantages, it is not a complete solution.

### 3.3 Ethical issues

While 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. It has been discussed by corpus linguists in forums such as the *Corpora* mailing list, and some authors have directly considered their work in relation to ethical issues, for example Hasund (1998), Sampson (2000) and Rock (2001). However, a number of major works on corpus linguistics, 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 there are existing ethical guidelines for gathering and using data that have been developed by professional linguists and these are used routinely by corpus linguists. For example, the British Association of 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. Nonetheless, a survey of the corpus literature reveals precious few examples of explicit discussion of ethical issues in corpus construction and use. Given that one or two corpora seem not to be wholly in line with ethical best practice, as will be shown, we might reasonably conclude that research ethics is an area that corpus linguistics should consider in more detail. Ethical issues in corpus linguistics can broadly
be divided into four main areas: ethical issues affecting respondents in a spoken corpus, ethical issues affecting corpus builders, ethical issues affecting corpus distributors and ethical issues affecting corpus users.

### 3.3.1 Ethics and respondents

What type of ethical issues affect respondents? The spoken part of the BNC provides a good example. When this data was collected, many people generously agreed to have their speech recorded in naturalistic settings (Crowdy 1995). The result is a very useful dataset indeed. However, the people carrying those tape recorders around had to give informed consent – they had to understand that whatever they said when that tape recorder was running might well eventually be used in a corpus that would be available to all who cared to use it. They were sacrificing their privacy.

Not only the privacy of what is said in their conversation, but also the privacy of personal information, may be sacrificed by a respondent. In particular, they may provide information about themselves to the corpus compiler which is useful in the generation of demographic metadata but is not to be incorporated into the metadata itself. In this case, keeping that data secure in perpetuity is an important responsibility that a corpus builder must discharge. For example, a speaker may reveal their occupation and workplace to allow their social class to be determined, on the express understanding that such information is not to be passed to a third party. Or they may give a detailed account of all the places they have ever lived, to assist in the classification of their dialect. Preserving the original information is important, so that the social categories assigned to the respondent can be validated when necessary – yet ensuring that access to this data is on a strictly need-to-know basis is just as important.

The ethical imperatives arising from the respondents’ sacrifice of privacy were clearly addressed by the builders of the BNC. But an equally pressing ethical issue relates to the people the respondents spoke to on tape – they also sacrificed their privacy, and it was therefore necessary for them to give their consent in an equally fully informed fashion. This consent was collected by the respondents themselves; so we see that in this case the process of data collection has actually imposed an ethical obligation on the respondents. This raises complex and not entirely soluble problems. Respondents are, of course, not themselves researchers and can hardly be expected to accept the full weight of the ethical obligations that rightly adhere to researchers. In fact, we would argue, when spoken data collection (unavoidably) imposes such obligations on respondents, it only makes the researchers’ responsibility of ethical oversight that much more critical. It is the researchers who must ensure and guarantee that good ethical practice is followed by each respondent – and, if there is any doubt about whether consent procedures have been followed fully and correctly by any respondent, the resulting data cannot ethically be included in a corpus.
Less obvious than the privacy of speakers in a corpus, but nonetheless important, is the privacy of the people talked about in the corpus. Their privacy is also arguably breached at times. Sampson (2000: section 4.1) summarises the problem well when he notes that in the BNC, two speakers:

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.

It would be much more difficult to obtain the consent of people who are not participants but are merely discussed in a given conversation. Accordingly, the privacy of all should be protected by anonymisation – names should be changed, though this may be done in a way which retains linguistically useful features of the name (e.g. a name typical of a female should be substituted with another name typical of a female). In addition to names, however, other personal data should be anonymised – there may be references to home addresses in a corpus, for example. The following example from file F86 of the BNC (utterances 264 and 265) shows how superficial anonymisation may be when sufficient information is left in the text to allow it to be circumvented. In this example, surnames have been deleted to provide anonymisation (as shown by the XML <gap> tags), yet the remaining context is clearly sufficient to allow them to be discovered.

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.

Though the example above is relatively benign, we can imagine serious consequences arising from a failure to take privacy seriously in the construction of a corpus. The Speech Act Annotated Corpus (Leech and Weisser 2003) contained credit card details that had to be anonymised, while the Lancaster Corpus of Children’s Writing (Smith et al. 1998) contained the personal details of young children. The example in Figure 3.1, from file HE7 in the BNC, shows clearly
<table>
<thead>
<tr>
<th>Speaker</th>
<th>Utterance Number</th>
<th>Utterance</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>275</td>
<td>Well don’t you think that it’s really rather improper for you to be doing this?</td>
</tr>
<tr>
<td></td>
<td>276</td>
<td>After all people are entitled to some secrecy aren’t they, about their confidentiality. You don’t feel that there’s any need at all to give any explanation of your behaviour?</td>
</tr>
<tr>
<td></td>
<td>277</td>
<td>You don’t think that an explanation is due here?</td>
</tr>
<tr>
<td></td>
<td>278</td>
<td>This information after all should have received confidential and does belong to other people, doesn’t it?</td>
</tr>
<tr>
<td>B</td>
<td>279</td>
<td>What I think is embarrassing is that doesn’t it?</td>
</tr>
<tr>
<td>A</td>
<td>280</td>
<td>And you’re just stealing it so that you can make money aren’t you?</td>
</tr>
<tr>
<td></td>
<td>281</td>
<td>Do people have a right to have their health records confidential do they not?</td>
</tr>
<tr>
<td></td>
<td>282</td>
<td>Have you got nothing to say whatsoever?</td>
</tr>
<tr>
<td>B</td>
<td>283</td>
<td>’Fraid not, no.</td>
</tr>
<tr>
<td>A</td>
<td>284</td>
<td>Robert is not alone in selling personal information from data banks.</td>
</tr>
</tbody>
</table>

Figure 3.1 A thief revealed.

why we may need to consider the privacy of those spoken about in a corpus. It would be problematic if, due to a corpus being published, a conversation like this led to a person being investigated by the police. In a case like this, if the data is not anonymised appropriately it might well be possible, going off what is said in the corpus, to walk around to the house of the person being accused of stealing, knock on their door and show them the transcript. Editing and anonymising such material in a transcript is one matter; ensuring that anonymisation occurs in audio and video recordings is obviously equally important, though much more challenging technically. It is now possible to construct mixed corpora of video, audio and textual material from a range of sources. But when the different media in such a mixed corpus are closely integrated, this can increase yet further the level of detail about specific individuals that is present, making anonymisation of the data yet more problematic; so it is not surprising that such corpora are rarely made publicly available.

The COLT corpus is a good example of a thoughtful approach to the issue of anonymity in corpus building, with due consideration being given to a range of ethical issues over time, as Hasund (1998: 16–17) discusses:

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.

In spite of this policy, the COLT team eventually decided, after further consider-
ation, that their approach to anonymisation was ethically and legally problematic
from the point of view of informed consent. They decided, in a later project
focused upon Norwegian teenage speech, to change the statement given to par-
ticipants to make it clear that their first name would be retained in the published
corpus (Hasund 1998: 24–5). This thoughtful, developmental approach to the
issues that both the law and ethics present to corpus linguists is clearly what is
needed in a field which is relatively new and which is using technologies and
methods which are relatively novel. There is, however, a price to be paid for giv-
ing due consideration to such ethical issues. The observer effect is undoubtedly
amplified by adopting an appropriately ethical stance to the gathering of sponta-
neous data of this sort. If the respondent knows that they are being recorded, the
chance that they will adapt their speech accordingly is undoubtedly elevated.

3.3.2 Ethics and corpus builders

Ethical issues also attach to corpus builders. We will use an example
from our own experience. 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), we were working very much in an opportunistic mode –
there was little available data covering all the languages we needed to include
in our corpus. We were approached by a religious organisation which wanted to
help; they translated many of their magazines and leaflets into the languages in
question from English originals. While this represented a golden opportunity to
expand our corpus, we felt we had to decline their offer. In part it was because
they saw the corpus, which is used in South Asia, as a way of distributing their
material and thus gaining converts. We were uncomfortable with the idea of
corpus work becoming missionary work. More importantly, when we surveyed
the material itself we found the material to be offensive in our view; for instance,
one magazine ran an article entitled ‘Who it is alright to hate’ (sic). Our aim
was not to construct a corpus of missionary texts. Nor was it to construct a
corpus of morally censorious texts. We certainly had no wish to be involved in
proselytising. Accordingly, we decided that the data compromised our ethical
stance and rejected the offer of the data. This example is at least clear. Our
dilemma might have been more acute if we had been working with a sampling
frame for which religious texts are required, such as the Brown Corpus sampling
frame. In that case, rejecting the material on grounds of offensiveness might
arguably have been unethical, since we would have effectively been skewing the
balance of the corpus towards particular philosophical or theological perspectives.
This is clearly more a question of research ethics – ethics relating to the conduct of scientific experiments – rather than a matter of the rights of respondents or text owners. It is a useful example to demonstrate the nuanced, multifaceted nature of ethics that corpus linguists should consider.

Returning to the example at hand, we did not feel we had an issue of research ethics in this case as we were aiming only to collect samples of official documents, very broadly defined, and so the issue of skew did not arise. In consequence the offensiveness of the texts became the primary ethical consideration. Any corpus builder may potentially find themselves in a position where they are forced to confront an issue like this, because 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.

3.3 Ethics and corpus distributors

Corpus distributors may face ethical issues of their own. For example, consider that you had built a corpus of texts exploring the language of a radical terrorist group. Would you be content to make your analyses available to the security forces of the country in which the group operated, to help them develop a counter-propaganda campaign? You might or you might not, but there are undoubtedly ethical issues to be considered, especially if you had worked with the radical group to gather the material on the understanding that the texts would not be given to third parties for whatever purpose. Real examples close to this one exist. It might be difficult to conceive of corpora being deployed in support of the US military, but in the USA, much corpus-based work is funded by DARPA, the Defense Advanced Research Agency. On its website is the statement ‘DARPA’s mission is to maintain the technological superiority of the U.S. military.’ Any publically available corpus may be used on DARPA-funded projects, quite legally and legitimately. But the technological superiority of the US military is not a goal that everyone on the planet would support. So here is a situation where a corpus made openly available could end up, without any legal obstacles, being used for purposes that the original corpus builders, and those contributing to the corpus, might never have agreed to. Personally, we do not disapprove of DARPA’s use of corpus data, and this is just one example out of many that we could have used. But what it exemplifies is that there are clearly ethical issues to be faced by corpus distributors when they are establishing the rules under which a corpus is redistributed. The offer of religious texts which we discussed above would have posed an ethical problem of a similar sort. Had we built a corpus knowing that a possible outcome of a user reading the material was religious conversion, we would have had to think very seriously about the issue of distribution. Quite apart from our personal ethical considerations, there are countries where seeking converts to a religion is widely regarded as immoral or is, in certain circumstances,
illegal. So at the very least, we would have had to address this issue by warning potential users of the nature and intent of the material when we distributed it.

Another key issue for corpus distributors, whether individuals or organisations, is the responsibility they bear for making sure that the data they hold remains intact and available. As a key goal of corpus linguistics is to aim for replicability of results, data creators have an important duty to discharge in ensuring that the data they produce is made available to analysts in the future. To simply delete data, or to deny it to future researchers, without a very good reason (for example ethical or commercial considerations of the kind that would necessarily take precedence) would be a great breach of trust and certainly represent a dilemma in terms of research ethics.

### 3.3.4 Ethics and corpus users

Finally, what ethical choices face the user of a corpus? Some may arise from the nature of the analyses. This can be shown very clearly in the case of forensic linguistics (Coulthard and Johnson 2007), an area which has, arguably, grown in power and significance because of the availability of corpus data and the techniques developed by corpus linguists (see McEnery et al. 2006: 116–20). In forensic linguistics the choices that the analyst makes are closely tied to a very serious outcome – somebody may be sent to prison unjustly, or a guilty person may walk free. Analysts have to think very carefully in such a situation about the consequences of the analysis they are undertaking, in particular how credible and reliable that analysis is. Normally a corpus analyst may be prepared to accept the possibility of error in an analysis; after all, such errors can be uncovered and corrected by repeating and refining the analysis over a period of time. But in forensic linguistics this approach to the data would be cavalier in the extreme, given the very high, very real stakes involved.

It should also be noted that corpus users have an important ethical duty to make the analyses on which their results were based available to future researchers, again in the interests of replicability. This point intersects with our discussion of corpus annotation (section 2.3). Making analyses clear and available to all is clearly good ethical and scientific practice. It has another dimension, however. Those analyses may be based on algorithms embedded in particular computer programs – if so, maintaining these programs, or at least maintaining a clear record of the detailed procedure by which they operated, may be as important as maintaining the results they produce. While this sounds straightforward in theory, in practice it can be quite difficult. Programs such as part-of-speech taggers are often continuously updated and improved by their developers. A corpus analysed by the CLAWS part-of-speech tagger (Garside et al. 1987) in 1989 will have a rather different pattern of errors to a similar corpus tagged by CLAWS six months ago – and unless careful records are kept of which version of a program has been used on each text, this information will be lost. It is entirely possible that, over
time, a researcher might produce (and, critically, publish) different sets of results that are based on the same data, but with different analyses produced by different versions of the same program. So there is a clear need to consider the storage and archiving of these analyses. In this case, ideally the program and its resources should be properly version-managed, and a record of every version archived, in the interests of replicability. The point does not simply apply to automated analyses, however. It applies to all analyses, including manual analyses of the corpus data. If a researcher decides that there are seventy-eight examples of a particular feature in a corpus, then it is incumbent upon that researcher to maintain a record of where and what those examples were. As we argued in Chapter 2, corpus annotation can be of help here as it allows the user to encode the analyses into the actual corpus data.

Another problem that a corpus user faces is that an analyst cannot be sure how their results will be interpreted by others. This is most serious when results are misinterpreted – and the misinterpretation very widely disseminated – by the mass media. The press is notoriously poor at accurately reporting on scientific research (see Goldacre 2008 for many examples). Of course, linguistics is not nearly so often the focus of press interest as the physical, and especially biomedical, sciences. But it would be foolish to imagine that corpus linguistic research is safe from media misinterpretation. For example, research into recent or current language change can often become confused with popular narratives of prescriptivism and language decay. Outside linguistics, it is widely believed that the ‘proper’ standards are being abandoned and language is becoming sloppier and less correct, a notion which can become interlinked with political anxieties over social decay (see Cameron 1995: 78–115). Press coverage of Leech’s research on the changing use of modal verbs (see section 5.3.2), for instance, often viewed the changes that Leech reported through this filter – which led in some cases to condemnation of Leech himself! In practice, there may not be much a researcher can do to correct mass media misinterpretation of their work. Regardless, we would argue that they have an ethical obligation to try – as a part of the more general responsibility of academia to the society that supports it.

3.3.5 Some cases of ethically problematic research

We have given, admittedly, no more than a thumbnail sketch of ethics in corpus linguistics. However, we may now consider how widespread good ethical practices have been in corpus linguistics to date. The literature, while somewhat silent on the question of ethics, generally embodies sound ethical practices. As noted already, corpus users do take privacy seriously when working with respondents. Corpus builders do not routinely produce corpora that are overtly unethical. Corpus distributors often go to great lengths to ensure that the data they distribute is produced to the highest legal and ethical standards – browsing the catalogues of the European Language Resources Association and
the Linguistic Data Consortium provides substantial reassurance to those who wish to consider ethical practices in corpus linguistics. Users of corpora often explore very sensitive issues in a nuanced and responsible manner; work in forensic linguistics is thorough and thoughtful. Linguists such as Baker (2008) explore sensitive issues, such as the representation of paedophiles, thoroughly and fairly. Nonetheless, it is possible to find instances of poor practice. The following examples could reasonably be argued to be in this category.

As noted, the BNC spoken corpus has a somewhat haphazard approach to anonymisation. Names are anonymised. Other features of the transcript which impact upon privacy are not. Worse, the original recordings that make up the BNC are available to the public and have not been anonymised at all – an afternoon spent listening to the recordings at the British Library would allow anyone to start unpicking the anonymisation that does exist in the text. This is not ideal. A response to this problem would be to restrict access to the recordings to ensure that this breach of ethics is limited (as happened with the COLT corpus, Hasund 1998: 16). This was the approach taken to an even more notable breach of ethics: as part of his work prior to the construction of the Survey of English Usage, Randolph Quirk gathered materials through surreptitious recordings (Quirk 1957). This is not necessarily a criticism of Quirk – by the standards of his time he was almost certainly not acting unethically. However, by modern standards this kind of data collection would definitely be judged unethical, and, as enticing as the prospect of using such data may be, it should clearly be shunned by linguists, and no such material should be gathered or used again. Finally, some data from the past is now simply lost – for example, while the work of Phillips (1989) is of great interest, it is quite impossible to replicate on the basis of the information in the published study; neither the algorithm nor the data used are fully presented.

In defence of most of the examples given, corpus linguistics has developed only recently, and the most marked examples of problematic ethical choices are firmly in the past. Some problems are attributable to the unstable nature of computer file storage in the late twentieth century; the ever-changing nature of file storage standards in that period quite understandably, though regrettably, led to a loss of data. With the spoken BNC, the researchers building it in the early 1990s were in terra incognita (just as Quirk was in the 1950s). A spoken corpus on the scale of the BNC had never been produced before. In doing something new, they made some mistakes. But the general message is that though corpus linguistics has made mistakes in the past, it has learnt from them, as noted. Nonetheless, ethical considerations remain an important area of development for corpus linguistics. We can only imagine that they will become ever more salient as the data available to the corpus builder grows, and the ability to cross-match it with other digital information sources, such as medical and educational records, emerges. As this capacity grows, however, so will the interaction between corpus linguistics and legislation – for instance, medical and educational records in the
UK and elsewhere are subject to legislative protection. The legal and ethical issues corpus builders have faced in gathering spontaneous speech, for example, may pale by comparison to those that they face when engaging with (meta)data that is subject to specific legislative protection.

### 3.4 Summary

This chapter concludes the two-chapter overview of what we might call the practical issues faced by corpus linguists. In the following chapters, we will explore different perspectives on corpus linguistics, looking at how they have been used in different traditions of corpus use. The next chapter surveys the work undertaken with corpora in the paradigm of English Corpus Linguistics. It was within that tradition that much of modern corpus linguistics developed, and that many of the issues discussed in this and the previous chapter were first explored.

### Further Reading

Hundt et al. (2007) is an indispensable collection of papers relating to the use of the web in corpus linguistics. The papers cover the philosophy of the ‘Web as Corpus’ approach but also sound some helpful notes of caution. In addition, the volume contains papers which demonstrate clearly how the web can be used very productively by corpus linguists.

For some criticisms of the Web as Corpus approach, see Leech (2007); for practical problems encountered by analysts when trying to ‘clean’ extraneous material from webpages for use in a corpus, see Baroni et al. (2008) and Hoffmann (2007a, 2007b) (and for a related problem, see the account on Jean Véronis’s blog of his heroic efforts to fathom the meaning of the ‘counts’ given by Google).

Less is available when we turn to the legal and ethical aspects of corpus use. Considering the importance of legal questions when constructing corpora, it is surprising how little has been written on the topic, with most books on corpus linguistics mentioning it in passing, if at all. McEnery et al. (2006: 77–9) has a brief section devoted to legal issues which is worth reading. Similarly the issue is addressed briefly by Hundt et al. (2007).

If little has been written directly on legal aspects of the use of corpus data, even less has been written on the topic of corpora and ethics, with all of the major works on corpus linguistics being silent on the issue. As noted in this chapter, individual papers have addressed the issue, but no overarching work looking at corpora and ethics has appeared. This is perhaps because, as noted, corpus linguists ‘inherit’ ethical guidelines and issues from other areas of linguistics, notably applied linguistics. To develop a sense of how corpus linguists do at times confront ethical issues, readers are encouraged to read both Rock (2001) and Hasund (1998).
Practical activity

(A3-1) 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?

Questions for discussion

(Q3-1) 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?

(Q3-2) When we analyse a discourse, we are also contributing to that discourse – especially, but not only, when our results are published. What ethical obligations does this place on an investigator, especially one working on sociologically sensitive discourses, such as representations of a minority group or hate speech?

(Q3-3) 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?