Unit 9 Copyright

9.1 Introduction

We noted in unit 8.2 that a major issue in building DIY corpora is copyright. While it is possible to use copyright-free material, such data is usually old and a corpus consisting entirely of such data is not useful if one wishes to study contemporary English, for example. Simply using copyrighted material in a corpus without the permission of the copyright holders may cause unnecessary trouble. In terms of purposes, corpora are typically of two types: for commercial purposes or for non-profit making academic research. It is clearly unethical and illegal to use the data of other copyright holders to make money solely for oneself. Builders of commercial corpora usually reach an agreement with copyright holders as to how the profit will be shared. Publishers as copyright holders are also usually willing to contribute their data to a corpus building project if they can benefit from the resulting corpus. However, the creation of commercial corpora is not our concern in this unit. Rather, we will focus on DIY corpora for use in non-profit making research.

9.2 Coping with copyright: warning and advice

You might think that you need not worry about copyright if you are not selling your corpus to make a profit. Sadly, this is not the case. Copyright holders may still take you to the court. They may, for example, suffer a loss of profit because your use of their material diminishes their ability to sell it: why buy a book when you can read it for free in a corpus (cf. also Amsler 2002)? Copyright issues in corpus building are complex and unavoidable. While they have been brought up periodically for discussion by corpus linguists, there is as yet no satisfactory solution to the issue of copyright in corpus building.

The situation is complicated further by variation in copyright law internationally. According to the copyright law of EU countries, the term of copyright for published works in which the author owns the copyright is the author’s life time plus 70 years. Under US law, the term of copyright is the author’s lifetime plus 50 years; but for works published before 1978, the copyright term is 75 years if the author renewed the copyright after 28 years.

One is able to make some use of copyrighted text without getting clearance, however. Under the convention of ‘fair dealing’ in copyright law, permission need not be sought for short extracts not exceeding 400 words from prose (or a total of 800 words in a series of extracts, none exceeding 300 words); a citation from a poem should not exceed 40 lines or one quarter of the poem. So one can resort to using small samples to build perfectly legal DIY corpora on the grounds of fair usage. But the sizes of such samples are so small as to jeopardize any claim of balance or representativeness. We maintain that the fair use doctrine as it applies to citations in published works should operate differently when it applies to corpus building so as to allow corpus builders to build corpora quickly and legally. Limited reproduction of copyrighted works, for instance, in chunks of 3,000 words or one third of the whole text (whichever is shorter) should be protected under fair use for non-profit making research and educational purposes. A position statement along these lines has been proposed by the corpus using community articulating the point of view that distributing minimal citations of copyrighted texts and allowing the public indirect access to privately held collections of copyrighted texts for statistical purposes are a necessary part of corpus linguistics research and should be inherently protected as fair use, particularly in non-profit making research contexts (see Cooper 2003). This aim is not a legal reality yet, however. It will undoubtedly take time for a balance between copyright and fair use for corpus building to develop.

So, what does one do about copyright? Our general advice is: if you are in doubt, seek permission. It is usually easier to obtain permission for samples than for full texts, and easier for smaller samples than for larger ones. If you show that you are acting in good faith, and only small samples will be used in non-profit making research, copyright holders are typically pleased to grant you permission. If some do refuse you remember it is their right to do so and move on to try other copyright holders until you have enough data.

It appears easier to seek copyright clearance for web pages on the Internet than for material collected from printed publications. It has been claimed (Spoor 1996: 67) that a vast majority of the documents published on the Internet are not protected by copyright, and that many authors of texts are happy to be able to reach as many people as possible. However, readers should bear in mind that this may not be the case. For example, Cornish (1999:141) argues that probably all material available on the Web is copyrighted, and that digital publications should be treated the same way as printed works.

9.3 Unit summary and looking ahead

Copyright law is generally formulated to prevent someone from making money from selling intellectual property belonging to other people. Unless you are making money using the intellectual property of other people, or you are somehow causing a loss of income to them, it is quite unlikely that copyright problems will arise when building a corpus. Yet copyright law is in its infancy. Different countries have different rules, and it has been argued that with reference to corpora and copyright there is very little which is obviously legal or illegal (cf. Kilgariff 2002). Our final word of advice is: proceed with caution.

Having discussed the most of the key concepts and practices in corpus linguistics in previous units, we will move on to the final unit of Section A to discuss the use of corpora in language studies.