

# Copyright and the Web as Corpus

Cecilia Hemming  
[cecilia@hemming.se](mailto:cecilia@hemming.se)  
GSLT  
Department of Languages  
University College of Skövde

Monica Lassi  
[monica.lassi@hb.se](mailto:monica.lassi@hb.se)  
GSLT  
Swedish School of Library  
and Information Science  
University College of Borås

## Introduction

The aim of this paper is to get clarity on copyright issues when using corpus data or creating corpora built on documents published on the Web. Legislations dealing with information technology have a tendency of becoming outdated fairly quickly. When it comes to e.g. the practical use of Internet it can be considerably different compared to what the legislators have taken into account. An example of this is that documents on the Web are used to analyze the language and the use of language. Corpus linguists are not as much interested in the content of documents, as in the language in which they are written. A researcher, for instance, can study the context in which certain words are used, the language used in chat rooms or news groups, etc. The prime interest of the researcher is in these cases the language or language features used, not the information conveyed. This information, though, can often be part of copyrighted material which makes it necessary to consider some legal aspects before using the data.

## What is a corpus?

Corpus (plural: corpora) is Latin for "body" and could in principle be used for any body of texts, utterances, or other specimens. In linguistics, though, the term tends to have a more specific meaning and a definition often found nowadays refers to a *machine-readable*, most often *finite-sized* body sampled in order to be maximally *representative* of the language variety under consideration. But such corpora have disadvantages: their size is limited, after a while they do no longer represent the current language, and their availability is restricted by their means of distribution (Volk, 2002). A recent trend is to use the Web for corpus linguistic tasks on various levels: lexicography, syntax, semantics and translations.

## What copyright issues have to be cleared?

When it comes to copying or publishing data it is very important to check up on copyright issues. Copyright laws handle in general cases where someone makes money from selling

intellectual property. If a corpus is sold, the copyright holders of e.g. texts within it easily react negatively towards the fact that someone earns money by selling their work (Kilgariff, 2002). Another issue, stated by (Amsler, 2002), is when copyright holder suffers a loss of profit because the use of their material diminishes their ability to sell it at full price.

### **Copyrighted material on Internet**

There seem to be some discrepancies in how much of published works on the Internet that are copyrighted. Spoor (1996, p. 67) claims that a vast majority of the documents published on the Web are not protected by copyright, and that many authors of texts are happy to be able to reach as many people as possible. In these cases copyright would not be an issue. This is a different view than that of (Cornish, 1999, p. 141) who claims that probably all material available on the Web is copyrighted, and that digital works therefore should be treated the same way as printed ones.

### **Copying or not copying**

One of the core prerogative of copyright is according to Spoor (1996, p. 67) the right of reproduction – copying. Copying is one of the most common operations on the Internet – files are copied from a server to the user’s hard disk or RAM, often with intermediate servers also copying the files before they reach the user’s computer. As stated by (Kilgariff, 2002), it is not at all clear that, for instance, downloading a report onto a PC-desktop for private use is any different to downloading it into a corpus for in-house use. Questions that arise in the light of this is: Can a digital document be considered as *copied* as soon as it is copied from a host server to intermediate servers, before ending up in the end user’s computer? Do the copies made on intermediate servers count as far as copyright is concerned? And does the copy in the user’s computer have to be stored on a disk or just in RAM to be considered copied?

Traditionally, there is an exemption from copyright when only short extracts are taken: brief quotes in reviews or academic writings seldom require special permissions. There are different opinions, though, about the significance of briefness in this regard. Poems, for instance, can contain very few words totally, so here the term “brief” should be interpreted as a proportion of the whole. (Kilgariff, 2002) advises corpora builders to avoid including anything, e.g. texts or parts of texts, where there is no explicit reason to do so. On the Web, there is a convention of leaving the right to authors to decide if their work is free to be viewed by robots or not, this view should according to Kilgariff also be transferred to the domain of corpus compilation.

### **Different nations, different laws**

Kilgariff stresses that before using material on the web one needs to get copyright clearance from all copyright holders regarded: authors, owners, publishers, all speakers for spoken material, etc. sometimes from many different countries (Kilgariff, 2002). The gathering of all these clearances can demand a lot of extra work and be very time consuming. When using/publishing texts on the Web there is not only one country’s law to worry about but there can be several. In the Corpora-list archives (Davies, 2002) claims that lawyers/professors, specialized in copyright law as it applies to the Internet, state that:

*the copyright law that matters is the law of the country from which the corpus materials are distributed, NOT the country where the original texts were created OR the country from which*

*end users access the materials.*

This, considering a corpus-project where material presented to the end user (short context concordance lines) was found radically different from the complete texts in the original format.

Torremans (2000, p. 106) describes two approaches to copyright issues. On the one hand the approach taken in continental Europe that, according to Torremans, concentrates on the rights that the creator of a work have, especially the economic rights. On the other hand there is the Anglo-Saxon approach, which can be described as concentrating on commercial exploitation of copyrighted work. Torremans claims that the time has come to integrate these both views, in order to handle the copyright issues globally. He proposes a model for harmonisation of different national laws in the European countries that combines rights of paternity, integrity and partial waiver. The paternity right is the right of an author to be recognized as the creator of a work. The integrity right is intended to protect authors from derogatory treatment of a work (additions, deletions or other changes to a work) and possibility to object to any changes made. The partial waiver gives the author a possibility to waive, or relinquish, the rights of the work, which can be the case if the work is to be commercially exploited by someone else than the author. This model is meant to protect the author's interest in a work and at the same time encourage honest commercial use of it (Torremans, 2000, p. 106-114).

### **Economic versus moral right**

Torremans discusses what happens to the moral rights of works when they are in digital form and available on the Internet. He claims that most copyrighted works are in fact published on the Internet. "Which law will apply to the issue of moral rights? And doesn't the very nature of moral rights require a uniform approach if these rights are to be protected effectively?" (Torremans, 2000, p. 99). When discussing copyright, the economic rights are most often the focus. For many authors, the moral right is as important though, or even more important. The problem is the same as with economic rights: no uniform approach exists as to who owns the moral rights to a work, and even what the scope of these rights are (Torremans 2000, p. 113).

Moral rights concern such aspects like that it is the author, and nobody else, that is attributed to a work, and that no changes should be made to it (Torremans, 2000, p. 99). This kind of right should, according to (Torremans, 2000, p. 103-105) be considered fundamental to protect authors against abuse of their work. He means that the law of the protecting country should apply not only to economic rights but also to moral rights. The protecting country is in this context the country in which a work is being used or the exploitation of a work is taking place. Torremans claims that "the application of a single harmonised choice of law rule would clearly constitute major progress in the digital online environment, but it must remain a second best solution". This would still result in different national rules on moral rights, which makes it very hard to handle global exploitation of copyrighted works (ibid, p. 105).

### **Myths about copyright on electronic texts**

National and international laws and rules on copyright issues in this area are still in its infancy, and it is hard to point out obviously legal or illegal standpoints. In the Corpora list

archives Brad Templeton (Templeton, 1995) gives an overview of frequent statements and misunderstandings:

**1) "If it doesn't have a copyright notice, it's not copyrighted."**

Today almost all major nations follow the Berne copyright convention which says that the default assumption for other people's works is that they are copyrighted and may not be copied unless one "know" otherwise. (Templeton, 1995) In many countries copyright of the text itself expires more or less 70 years after the death of the author. This is the case in countries that are members in The European Economic Area (ERA), in May 1997: Austria, Belgium, Denmark, Finland, France, Germany, Greece, Iceland, Ireland, Italy, Luxembourg, the Netherlands, Norway, Portugal, Spain, Sweden and the UK (Cornish, 1997, p. 28).

**2) "If I don't charge for it, it's not a violation."**

This is not true. The copyright is violated whether you charge or not, but as the commercial value of the property can be hurt, it can affect the damages awarded in court. (Templeton, 1995)

**3) "If it's posted to Usenet it's in the public domain."**

False, nothing is in the public domain unless the owner explicitly puts it in the public domain by saying something similar to "I grant this to the public domain". Usenet is no exception. Spoor (1996, p. 67) points out, though, that some works are distributed on the Web because the creator wants the work to be distributed as much as possible, i.e. news group discussions.

**4) "My posting was just fair use!"**

Fair use is almost always a short excerpt and almost always attributed. Reproduction of an entire work is generally forbidden. If the right to comment overrides the copyright is up to the court to decide.

**5) "If you don't defend your copyright you lose it."**

Nowadays, copyright is in effect never lost, unless it is explicitly given away.

**6) "Somebody has that name copyrighted!"**

A correspondence to copyright when it comes to short things like names or titles is *trademark*. You can only trademark a word in a special context, and owning a mark does not mean complete control.

**7) "They can't get me, defendants in court have powerful rights!"**

If you violate copyright you usually get sued, not charged with a crime. "Innocent until proven guilty" is a principle of criminal law, as is "proof beyond a reasonable doubt."

**9) "It doesn't hurt anybody -- in fact it's free advertising."**

Only the owner can decide if he wants free ads or not.

**10) "They e-mailed me a copy, so I can post it."**

To have a copy is not to have the copyright.

## Practical cases

### The European Language Resources Association (ELRA)

ELRA has adopted a way of handling the copyright problems by acting in between the two parties. This, to encourage the sharing of data, originally developed for specific language resource provider's internal needs, and to simplify the relationship between the users and the providers/producers of the resources. First, a distribution agreement is set up between ELRA and regarded copyright owners or providers. Then a contract, defining the responsibilities and obligations of each end user or value added reseller (VAR), are established between ELRA and the respective user/VAR. See figure X for a brief overview of this process.

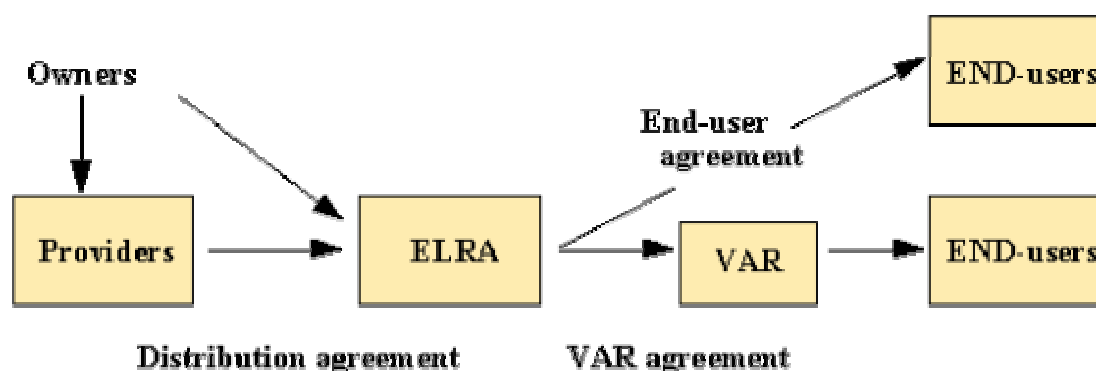


Figure X. (<http://www.icp.grenet.fr/ELRA/services/legals.php#princip>)

There are a number of corpora that include Web documents and that are available directly on the Web. Three of these are presented below, emphasizing how the creators handle the copyright issues.

### Korpus 2000<sup>1</sup>

Korpus 2000 is a both a project and a product (a corpus) of the Danish language. The corpus contains about 25 million words collected from about 20.000 different texts from 1998-2002. The texts are of many different types, both private and public – books, journal articles, newspapers, advertisements, letters, diaries etc. The purpose of the corpus is research in i.e. linguistics and stylistics of text. Therefore full texts will not be available, but only a small portion to demonstrate adjacent context. The focus here is the use of language in the texts and not the texts as such. Another reason why entire texts are not published by Korpus 2000 is to protect the project from possible copyright infringement in case they do not have the rights to publish a specific work. (Korpus 2000, 2002)

---

<sup>1</sup> <http://korpus.dsl.dk>

## **The Oxford Text Archive (OTA)<sup>2</sup>**

The OTA was created over twenty years ago and claims to be one of the largest electronic archives. It aims at providing well documented texts for research and educational purposes (Reid 2002c), which means that the texts in the corpus are not to be sold or published unless the creator of the text has given explicit permission (Reid 2002b). If these rights are to be given directly to the user or to the OTA is unclear. The OTA claims that the rights to the works included in the corpus are either outdated, or cleared with the authors. On the Web site, however, there is a plead to the users of the corpus to contact the organisation immediately if they happen to see any work that is not cleared with the copyright owner (Reid 2002a).

## **WebCorp<sup>3</sup>**

WebCorp is created at the University of Liverpool and offers corpus linguists and other people interested in language a set of tools which allows access to the Web as a corpus. WebCorp uses the output (in the form of URLs) of some search engines to extract concordance lines for the words searched (the user can pick one search engine at a time) (Research and Development Unit for English Studies, University of Liverpool 2002). Considering the tradition of copyright exception for short extracts mentioned above there would be no copyright issues to consider: WebCorp does neither store nor publish the documents that are used to create the concordance lines.

## **Conclusions**

The general assumption among the sources cited in this paper is that copyright issues of digital works should be treated the same way as they are for printed works. There are important divergences though. An issue that has to be handled somehow is that, in contrast to what is the case for printed works, national laws in this matter are hard to follow. There is a lack of uniformity in what is protected by the laws, and how the laws in different nations work. Some countries put the created work in focus, while others focus on the creator. The moral right of a work, fundamental to protect creators against abuse of it, is also a discussed question and just as with economic rights there is no uniform approach to it.

The intended use of the corpus to be created is an important aspect of a would-be copyright problem. If the intention is to use a corpus in-house for research purposes the copyright issue is a smaller problem than if the intention is to make the corpus, and more importantly the works in the corpus, available for public use. The latter case can be seen as similar to publishing the individual works, which would necessitate permissions of all the copyright holders from respective nations. Traditionally, there is an exemption from copyright requests when only short extracts and brief quotes are presented; an example of this is the concordance line output of WebCorp. A problem though could be the interpretation of “short” and “brief”.

Not many of the copyright issues that arise by the use of the Web as a corpus base, and that have to be resolved, have actually been resolved. This is also why misunderstandings flourish. Considering Torremans’ claim, that most copyrighted works are published on the Internet (Torremans, 2000) and Templeton’s claim that works are presumed to be copyrighted until

---

<sup>2</sup> <http://ota.ahds.ac.uk/>

<sup>3</sup> <http://www.webcorp.org.uk>

proven otherwise (Templeton, 1995), the creator of a corpus has a great deal of work to do in controlling the copyright, and in some cases getting clearance or clearances to use them. One solution can be represented by the ELRA approach: they put themselves in between the resource provider and user, thereby simplifying their relationship and the sharing/reuse of data.

## References

- Amsler, Robert. 2002. In *Corpora List Archive*. Legal aspects of corpora compiling. <http://helmer.hit.uib.no/corpora/2002-3/0256.html> 2002-11-04
- Cornish, G. P. 1999. *Copyright: Interpreting the law for libraries, archives and information services*. Library Association Publishing, 3. ed.
- Davies, M. 2002. In *Corpora List Archive*. Legal aspects of corpora compiling. <http://helmer.hit.uib.no/corpora/2002-4/0017.html> 2002-11-04
- Kilgariff, A. 2002. In *Corpora List Archive*. Legal aspects of corpora compiling. <http://helmer.hit.uib.no/corpora/2002-3/0253.html> 2002-11-04
- Korpus 2000. 2002. <http://korpus.dsl.dk/korpus2000/beskrivelse.php> 2002-11-13
- Reid, A. 2002a. *Collection Policy*. [http://ota.ahds.ac.uk/publications/ID\\_AHDS-Publications-Collections-Policy.html](http://ota.ahds.ac.uk/publications/ID_AHDS-Publications-Collections-Policy.html) 2002-11-27
- Reid, A. 2002b. [http://ota.ahds.ac.uk/faq/section\\_DEPOS\\_1\\_g.html#1\\_g](http://ota.ahds.ac.uk/faq/section_DEPOS_1_g.html#1_g) 2002-11-27
- Reid, A. 2002c. [http://ota.ahds.ac.uk/faq/section\\_UFAQ\\_1\\_g.html#1\\_g](http://ota.ahds.ac.uk/faq/section_UFAQ_1_g.html#1_g) 2002-11-27
- Research and Development Unit for English Studies. 2002. *WebCorp Frequently Asked Questions*. University of Liverpool. <http://www.webcorp.org.uk/webcorp.html> 2002-11-13
- Spoor, J. H. 1996. The Copyright Approach to Copying on the Internet: (Over)Stretching the Reproduction Right? in *The Future of Copyright in a Digital Environment* (ed. Hugenholtz, P. B. ). Kluwer Law International.
- Templeton, B. 1995 <http://www.clari.net/brad/copymyths.html> 2002-11-10
- Torremans, P. 2000. Moral Rights in the Digital Age. In *Copyright in the New Digital Environment: The Need to Redesign Copyright* (eds. Satamatoude, I. A. and Torremans, P. L. C.), Sweet & Maxwell.
- Volk, M. 2002. Using the Web as Corpus for Linguistic Research. In *Tähendusepiudja. Catcher of the Meaning. A Festschrift for Professor Haldu Õim*. (eds. Pajusalu, R. and Hennoste, T.) Department of General Linguistics 3, University of Tartu.