Open access to sign language corpora

Onno Crasborn
Department of Linguistics, Radboud University Nijmegen
PO Box 9103, NL-6500 HD Nijmegen, The Netherlands
E-mail: o.crasborn@let.ru.nl

Abstract

This paper sketches recent developments in internet publishing and related copyright issues, and explores how these apply to sign language corpora. As a case study, the *Corpus NGT* project is characterised, which publishes a systematic collection of sign language video recordings and annotations online as open access data. It uses Creative Commons licenses to make explicit the restricted copyright rules that apply to it.

1. Background

While native intuitions of Deaf informants have played some role in the linguistic study of signed languages, linguistic studies since Tervoort (1953) and Stokoe (1960) have mostly used film and video recordings. Descriptions and transcriptions of these video recordings were made on paper until the 1980s; since then, the transcriptions were increasingly made in office software like word processors, spreadsheets and databases. It was not until the 1990s that digital video became commonplace, and only since around the year 2000 it has become easy to process and store large amounts of video recordings in desktop computers. Only since the venue of multimedia annotation tools like SignStream, Transana and ELAN sign language researchers can use a direct link between their transcriptions and video recordings.

This paper will not go into the technical aspects of these developments, but aims to describe the ongoing shift in accessibility to sign language data by researchers. Many sign language researchers and research groups used to have shelves full of video tapes, but were not able to use the data very often after an initial transcription or analysis was made, simply because of the extremely time-consuming process of locating a specific point in time on a video tape, let alone comparing different signers on different tapes. With the use of modern technology, a direct link can be established between an instance of a transcription value and a time segment in a particular video file, and data that are already transcribed can easily be double-checked or shown to colleagues. This is commonly seen as leading to a potential increase in quality of one's own research.

We are currently at the brink of a next step in our use of sign language data, as data can be exchanged over internet and even published online. In this way, it can become easier to also check data used for linguistic publications by other investigators; access to not only the linguistic analysis but also the data at the base of that analysis could lead to a further increase in reliability of linguistic research. This may appear to be obvious, as linguistic analysis typically do include written examples of the data under discussion for languages like English or Spanish, or phonetically transcribed examples of unwritten languages. The situation is a bit different for signed languages, where there is no conventional writing system that in use throughout deaf communities, and moreover, there is very little standardisation on the transcription of sign language data, whether for gloss annotations or for phonetic transcriptions. This holds both for manual and non-manual activity.[1] Access to the original data therefore has a relatively large value in evaluating linguistic claims.

Aside from the technological difficulty in creating digital video files, there are privacy issues related to the publication of video material, as not only what is said can be accessed, but also (and more unequivocally so than with audio data of speakers) what the identity of the speaker or signer is. This paper describes the ongoing developments in the publication of data on internet (section 2), and then discusses the nature and role of privacy protection of online publications of video recordings (section 3). Section 4 characterises one particular system of user licenses that is being developed especially for online publications, 'Creative Commons'. As a case study, section 5 discusses the construction and open access publication of the *Corpus NGT*, a linguistic corpus of video recordings of Sign Language of the

[1] SignWriting (http://www.signwriting.org) is used as a writing system in parts of some deaf communities and HamNoSys (http://www.sign-lang.uni-hamburg.de/hamnosys) and FACS (http://face-and-emotion.com/dataface/facs/new_version.jsp) are stable phonetic annotation systems, but as yet, none of them is actually used by a substantial part of the research community.

Netherlands, which makes use of Creative Commons licenses to protect the data from undesired types of use.

2. Internet publishing developments

The publication of speech resources for spoken language research is quite common, and text data have been an object of study since the earliest stage of computer technology. There are now several organisations that offer online speech resources and associated tools for sale, including the Linguistic Data Consortium (LDC)[2] and the Evaluations and Language resources Distribution Agency (ELDA)[3]. Increasingly, spoken language data are also recorded and published on video, to be able to study non-verbal behaviour of speakers in addition to speech. The organisations above typically sell copies of data sets to researchers, rather than simply publishing them on a server for everyone to access for free. The intent is not necessarily to make profit from these sales; sometimes, the goal is merely to cover the costs that are made in creating hardcopies of data and manuals and sending them to someone.

One of the current developments on internet more generally is the increasing attention for 'open content': data of all kinds, whether text, images or video, are made publicly available, without charging a fee. While there may be restrictions on the type of use that is allowed, selling content and strictly protecting it under copyright laws appears not desirable necessary for some types of content. For example, many (starting) artists benefit from the wide distribution of their creative output without wanting to sell specific instances of works of art. For new art forms that crucially depend on computer access, including some multimedia productions, free internet access is a crucial component of their work. In addition to audiovisual and graphic arts, text distribution can also profit from open access even though traditionally, essays would be published in journals or books that could only be obtained by purchasing them.

Traditional publications of reproducible work in hardcopy, whether on paper, CD or DVD, or any other medium, would typically be accompanied by a message stating that "all rights are reserved". When computer technology made the copying of for example music purchased on a CD easier, this statement did not so much apply to the unauthorised copying of parts of a text in another text, but to creating actual copies of the material. The venue of digital information distribution over internet was accompanied by new means of protection, referred to as 'digital rights management' (DRM).

By contrast to these commercial publications, there are now many publications on internet where the explicit goal of the author is not to prohibit copying and usage, but rather to encourage use by others. This development is sometimes characterised as a change from 'copyright' to 'copyleft': rather than stating that "all rights are prohibited", people are encouraged to use materials for their own benefit.

The same change in perspective can also be witnessed in science. Rather than being protective of one's own data, it is becoming more and more common to publish research data, hoping that others will profit from it and do the same with their own data. The European Research Council, founded in 2006, explicitly encourages open access to research data, noting that while hundreds of repositories exist for the medical and natural sciences, the humanities are in a different position:

> "With few exceptions, the social sciences & humanities (SSH) do not yet have the benefit of public central repositories for their recent journal publications. The importance of open access to primary data, old manuscripts, collections and archives is even more acute for SSH. In the social sciences many primary or secondary data, such as social survey data and statistical data, exist in the public domain, but usually at national level. In the case of the humanities, open access to primary sources (such as archives, manuscripts and collections) is often hindered by private (or even public or nation-state) ownership which permits access either on a highly selective basis or not at all." (ERC, 2006)

'Open access' does not necessarily imply that no restrictions apply, nor that anyone can view materials without registration or subscription; thus, in the area of science, archive access may well be restricted to people who register as researchers or who work at research institutes. The Creative Commons licenses discussed in section 4 constitute one way of restricting the use of materials, but imply no assumption on whether one needs to register to use the materials.

3. Ethical concerns in the publication of sign language data

As was already indicated above, the publication of sign language data on video implies inevitably that the message content can connected to the identity of the signer. Even without explicitly adding the name or other details of the signer's identity to the video clip in metadata, people can easily be identified on the basis of their face. The chance that this will happen as well as its potential consequences are relatively large given the small size of Deaf communities in most countries. For

[2] http//www.ldc.upenn.edu
[3] http://www.elda.org

example, in the case of the Auslan corpus that is currently being constructed at Macquairie University, Sydney, the 100 people in the corpus form 1.7% of the Australian Deaf community, estimated to be about 6,000 (Johnston 2004).[4]

The open access publication of a sign language corpus implies providing information on who is and who is not recorded for scientific data, which in such a small community can be a sensitive matter in itself. The wide range of possible uses of a corpus of a substantial subset of signers might also have an influence of the language, the signing in the corpus being considered a standard of some form, or the signers being considered role models for second language learners. These type of issues will not be further discussed here, but they are considered as meriting further attention in any corpus construction project and any publication of sign language data.

The recording of signers for any linguistic research typically does not involve special ethical reviews for dealing with human subjects, which are common in (international) grant applications: there is no risk of (physical or psychological) harm to the signer, participation is voluntary and signers typically receive payment for their contribution, they just need to be treated with respect. Moreover, people typically sign a form to give the researcher a proof of their 'informed consent', which means that (1) the person has the legal capacity to give consent (so that parents should give consent for participation of their children), (2) the person gives consent on a voluntary basis, not being pressured to participate, and (3) the person is able to make an informed decision. It is exactly this last point that warrants some further attention.

Firstly, depending on the type of data that are being recorded and published, a lot of personal information can be revealed in discussions and conversations. While it is attractive to use free conversation data as instances of spontaneous language use, the risk of including personal information (whether about oneself or about others) increases, and it is not always possible to monitor this before publication of the material, neither by the signer nor by the researcher.

A document guidelines for research ethics of linguistic studies from McGill University (Canada) characterises most linguistic data collection as being 'low-risk' in the sense that "the information being collected is not of a sensitive or potentially private nature, i.e. people would not reasonably be embarrassed by other people knowing about it" (McGill 2008). The problem with online publication of sign language videos is thus that the nature of the data cannot always be well established, but moreover, that publication on internet cannot be undone. While a publisher can in principle try to withdraw a publication by finding back all copies of books or CDs, this is virtually impossible with electronic open access material once it has been downloaded or re-distributed by others.

The irrevocable nature of the publication of sign language video data could also become a problem when signers decide in the future to withdraw their participation. Although the consent form has given the researcher the legal right to publish the material, for a good relation with the participant and the Deaf community in more general terms, it could be desirable to indeed withdraw items from a corpus that is already published.

Secondly, it is debatable whether anyone can make an informed decision on publication of video recordings on internet given the high speed of the development of computer technology. As publication entails possible availability forever, new technologies can imply uses of the video data that we cannot yet foresee. Although one can decide to not use names or initials in any of the metadata accompanying videos (as was done in the Corpus NGT, see section 5), if face recognition software should become available as part of the average desktop operating system and when automatic sign recognition technology allows translation of signed discussions to text (in whatever language), discussion content and identity can easily be matched and linked to further information on individuals that is available online. Thus, even though at present signers may be perfectly happy with the publication of video recordings, it is not unlikely that this will change in the future.

On the other hand, we currently also see a rapid change in what is considered as privacy-sensitive information now that people massively publish their own materials online. Aside from discussions in message boards and mailing lists, many people do not hesitate to publish large sets of family pictures online, and community web sites like Facebook[5] or Hyves[6] elicit wide participation from people who appear to be eager to share a lot of personal information with the whole world.

The question remains whether this is a sign of a permanent change in (Western) culture, or whether people will be dissatisfied with it in ten or twenty years time. Where people voluntarily take part in the publication of personal information about themselves, one might expect that this is not so much an issue, although one may still debate whether anyone can estimate the impact of exposing details of one's private life online. However, in the case of sign language corpus construction and open access publication, the decision to publish something online is very

[4] http://www.ling.mq.edu.au/centres/sling/research.htm

[5] http://www.facebook.com

[6] http://www.hyves.net

indirect: it is not a concrete activity of a signer at his own computer, but the signing that was recorded was not inspected by the signers, and was only published online a few months after the event. It will remain important to monitor and discuss these developments in the future.

4. Creative Commons licenses

Although copyright law cannot completely prevent abuse of published material, it can encourage people to treat materials with respect. Creative Commons is a recent initiative that explicitly aims to allow publishers of online material to apply some restrictions to the (re)use of online content, by declaring the applicability of a license with one or more conditions to a specific work that is published online. The international organisation Creative Commons was founded in 2001 as a bridge between national copyright laws and open content material on internet. All licenses have been translated to the national languages of more than thirty countries and have been adapted where necessary to national copyright laws in these countries, yet they all seek to stay as close as possible to the US originals to ensure that the licenses will be regarded as an international standard.

There are currently three types of restrictions, and some new developments are underway. The first restriction that can be applied is dubbed "BY", and requires the user to refer to the original author of the work when re-publishing or using the work. The second restriction concerns the prohibition of commercial use of the work, and is dubbed "NC" (no commercial use). The third restriction concerns the modification of the work, and states that the work has to be reproduced in the same form ("ND", no derivative works) or that modifications are allowed but have to be shared under the same conditions ("SA", share alike).

The Creative Commons licenses are available in various forms: a plain language statement (as in the previous sentences), a formal legal text, and a machine-readable version for use by software. Reference to the licenses on internet is typically done by including an images with symbols for the different license conditions, some of which ar illustrated in Figure 1. The image then links to the text of the actual license, or explicit reference to the URL of the license text can be included.

A large advantage of using these licenses is that creators of any type of work can publish materials themselves, and enter in an agreement with the user about the types of use that are allowed. Traditionally, various types of publishers acquired the rights for distribution, promotion, sales, et cetera, and these publishers then entered into agreements with the end users (here too, the term 'license' was sometimes used). Thus, using the Creative Commons licenses, creators can retain more responsibility over what happens to their material, and at the same time profit from the relatively cheep production and distribution channels that are now offered on internet. All rights remain with the creator of a work.

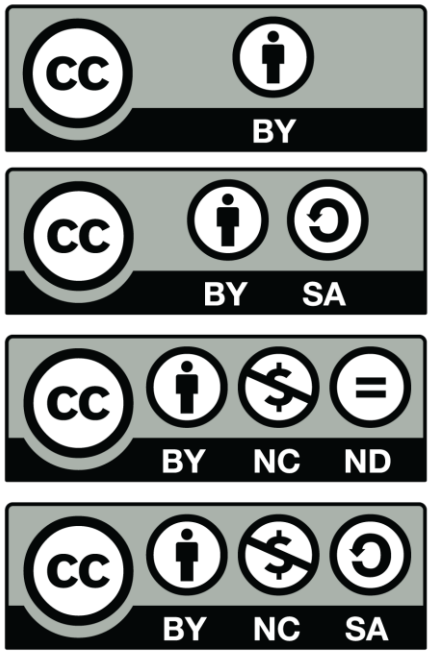

Figure 1. Examples of Creative Commons license buttons

5. A case study: the *Corpus NGT*

The Creative Commons licenses form a very attractive way of protecting the use of the sign language videos in the *Corpus NGT*, a sign language corpus of Sign Language of the Netherlands (Nederlandse Gebarentaal, NGT; Crasborn & Zwitserlood, this volume).

For this corpus, a total of 100 signers will be recorded; most of these will be available in the first release in May 2008. These signers produced around 75 hours of interactive language material, divided in more than 2,000 video segments. The wish to publish this material not only for research purposes (its primary goal, cf. the funding from the Netherlands Organisation for Scientific Research) stems from its large possible value for various parties in the Netherlands: deaf signers themselves, second language learners of sign language, interpreting students, etc.

As was discussed above, a central problem in publishing sign language data online is privacy protection. In the *Corpus NGT*, we try to protect the privacy of the informants in several ways: we urge people to not reveal too much personal information about themselves or about others in

their stories and discussions, we limit the amount of metadata that we publish online (leaving out many of the standard fields from the IMDI metadata standard), and nowhere we mention or refer to the name or the initials of the signers. Personal information about family background and signing experience that we did collect will in principle be made available for other researchers, who will have to sign a license form declaring not to publish information on individuals. The nature of this license is not yet established, but we might consider copying such agreements from endangered languages documentation projects such as DOBES.[7]

We chose to apply the Creative Commons 'BY-NC-SA' license to all of the movie files in the *Corpus NGT* (symbolised by the last image in Figure 1). This license states that people may re-use the material provided they refer to the authors, that no commercial use be made, and that (modifications of) the material are distributed under the same conditions. As opposed to the 'no derivative works' condition, the latter condition allows users to use segments of clips for their own web sites, to add subtitling or other graphics to it, et cetera. While these types of modification will not frequently be interesting to scientific users, they do broaden the possible educational uses of the material.

Although the permission for the licensed open access publication is requested of the signers in the corpus, it was discussed above that we can not guarantee that signers can foresee the consequences at the time of recording. Will future technologies allow easy face recognition on the basis of movies and thereby obliterate the privacy protection measures that have been taken? What will the (normative) effect of publishing signing of a group of 100 signers from a small community be? There is a clear risk in the publication of sign language data without an answer to these questions. The 'solution' taken in the *Corpus NGT* project is to invest substantial time and energy in publicity within the deaf community, to explain the goal and nature of the corpus online, and to encourage use by deaf people.

The plain language version of the licenses is attached to every movie in the *Corpus NGT* by a short text preceding and following every movie file, thus allowing relatively easy replacement should future changes in policy require so (Figure 2). We expect to offer a signed version of the licenses in the near future as well.

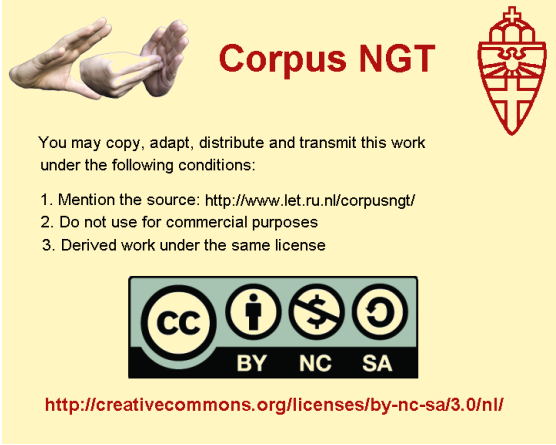

Figure 2. Reference to Creative Commons licenses in the *Corpus NGT* movies

6. Conclusion

The possibilities offered by current internet and video technologies together with new forms of licensing agreements offer attractive possibilities for the archiving of sign language research material, at the same time offering access to these materials for the language community itself and other interested public parties. This paper has tried to emphasise that the possibilities also raise new ethical issues that should receive attention at the same time. The traditional research ethics of informed consent and respecting ones informants will not be sufficient for internet publishing. The recently founded Sign Language Linguistics Society,[8] which is currently setting up a code of conduct for sign language research, might play a role in the discussion of these developments.

Acknowledgements

This paper was written as part of the Corpus NGT project and is supported by grant 380-70-008 from the Netherlands Organisation for Scientific Research (NWO).

References

ERC (2006) ERC Scientific Council Statement on Open Access. http://erc.europa.eu/pdf/open-access.pdf. Document accessed in March 2008.

Crasborn, O. & Zwitserlood, I. (this volume)

Johnston, T. (2004) W(h)ither the deaf community? Population, genetics and the future of Auslan (Australian Sign Language). American Annals of the Deaf, 148(5), pp. 358-375.

McGill University (2008) Department of Linguistics Procedures for Ethical Review of Research on Human Subjects.

[7] http:/www.mpi.nl/DOBES

[8] http://www.slls.eu

http://www.mcgill.ca/files/linguistics/research_ethics.pdf. Document accessed in March 2008.

Stokoe, W. (1960) Sign language structure. An outline of the visual communication systems of the American Deaf (1993 Reprint ed.). Silver Spring, MD: Linstok Press.

Tervoort, B. (1953) Structurele analyse van visueel taalgebruik binnen een groep dove kinderen. Amsterdam: Noord-Hollandsche Uitgevers Maatschappij.