ONNO CRASBORN

What Does "Informed Consent" Mean in the Internet Age?

Publishing Sign Language Corpora as Open Content

Between 2006 and 2008 a video corpus of Sign Language of the Netherlands (Nederlandse Gebarentaal, or NGT) was created with the support of the Netherlands Organisation for Scientific Research (NWO, grant no. 380-70-008).[1] While the original goal of the project was to create a large research database for linguistic investigation irrespective of the researcher's location and institution, early on in the project it was decided to make the data publicly available. Before this time, various parties in the Netherlands, including interpreter trainers, sign language teachers, and interpreters, had expressed considerable interest in such data. Given the absence of written resources for signed languages, the availability of video materials can potentially have a significant impact on deaf communities.

All ninety-two participants who were recorded for the project through December 2008 signed a consent form indicating they agreed to online publication. This article raises several issues relating to "informed consent" as it applies to the publication of sign language data as open content on the Internet. First of all, to what extent are deaf

Onno Crasborn is associate professor in the Linguistics Deparment at Radboud University Nijmegen and head of the sign language theme in the local research institute Centre for Language Studies.

people with varying levels of Dutch literacy aware of the status and impact of a consent form? Although the statements on the form were explained to them in sign language, one may wonder to what extent this counts as a voluntary and well-informed decision. Second, one may wonder whether it is possible to agree to the online publication of such recordings given the rapid technological developments that we have seen in the last decade. Just as few people would have foreseen the significance of sharing social data in applications like Facebook and Google Earth, we cannot predict the impact of new technologies. Will face recognition on the basis of movies be built in to every operating system in ten years' time? These are new types of considerations that all touch upon the "well-informed decision" that is inherent in informed consent. This article describes some current developments in this area on the Internet. The next two sections focus on new licenses to protect the use of data, and the section that follows them addresses the central question of the value of informed consent in the publication of sign language corpora.

Technological Advances in the Study of Signed Languages

The linguistic study of spoken languages has long been restricted to the analysis of written resources. For centuries, grammars and dictionaries have been based on written rather than spoken language. Text documents have been increasingly available and accessible, and the first computer technologies in the 1960s and 1970s were able to process only text, not audio or video recordings. In fact, it took quite some time before corpus linguistics as a separate branch of linguistics arose. Aside from technological impediments, it was not until Labov's observations of variation in speech in the 1960s that the study of the *use* of language (rather than the knowledge or structure of language) became an independent area of study. The rise of a separate discipline of corpus linguistics, which uses larger collections of texts as data, followed in the 1970s. The development of language technologies such as automatic speech recognition, which facilitated the study of speech corpora, did not take off until less than twenty years ago; this new phase enabled the study of speech behavior in addition to written texts, thereby allowing for insights into speech and everyday spoken

interactions, which typically constitute more of spontaneous language behavior than writing. With the rapid rise of Internet use and the mass publication of text on web pages, text corpora have become more prominent in linguistics as they constitute a rich source of information on everyday language use now available in huge quantities online.

By contrast, there is a dearth of resources of any kind for signed languages. Written materials have never really played a role in any deaf community. Notwithstanding recent efforts to promote SignWriting in education, very few people are "literate" in a signed language. More important, there has not been a written culture in the past decades, let alone centuries: Signed texts are not available in education or for leisure and by extension not for linguistic study, either. In fact, one might argue that the absence of a written culture makes it difficult to reflect on language. Such metalinguistic skills are based on linguistic intuitions that lead to knowledge of a language rather than the study of its use. However, the latter has been difficult because of the technological state of affairs as well. Playback of video recordings on personal computers has become possible only in the past fifteen years, and then only gradually. It was this development that caused a major breakthrough in the way that sign language data could be viewed: One could access and compare random fragments of a recording without having to endlessly play films or videotapes backward and forward. A crucial, related development is the creation of annotation programs such as ELAN (*E*UDICO *L*inguistic *An*notator), which allow users to tag aspects of signing in video recordings so that they can later be searched and reinspected. This is leading to the emergence of a field of "corpus sign linguistics"—without the use of sign language writing. Concomitant problems in the use of the written form of spoken languages to annotate sign language have been under discussion for quite some time (Johnston 1991; various articles in Bergman et al. 2001; Crasborn et al. 2008).[2]

Nowadays, even the simplest personal computer can store many hours of video recordings. Moreover, it has become common to view and share video files online. While the currently most popular web site, YouTube, may well disappear within a few years, the new concept of publishing will not. Any individual can publish a video recording and

make it accessible to the whole world. These and related technological developments that are under way are perhaps relatively small compared to the large step from analogue film and video to digital computer files, yet they may lead to a much greater revolution in the linguistic study of signed languages. In sharp contrast to this history of sign language data, there is now a medium that allows for a sign language parallel to the written culture that we know from many spoken languages. Online sign language videos can constitute new sources of data for some types of studies.

From the perspective of Deaf communities, there is a different prominent aspect to these developments. The rise of online video in itself is not a surprising event. It fits with the general trend that started with modern communication technology in the nineteenth century, which has led to major increases in mass communication. For Deaf people, however, who have missed the impact of telephone conversations and realized little benefit from national television in terms of long-distance communication, the rise of video on the Internet brings a condensed version of these technological advances, which have been spread out over several decades for hearing people. For that reason, the online publication of collections of sign language video material may have a greater impact on Deaf communities than may be foreseen.

For many linguistic studies, the use of random sets of language use by unsystematic collections of signers will not suffice. Often, more controlled data are needed, where at least the context of language use is known. Typically, such corpora strive for a balanced collection of signers in terms of different sociolinguistic variables such as age and regional background. The creation of sign language corpora, similar to those for text and speech, can address this need. Although a number of linguistic research projects have led to substantial data collections in the past, a linguistic corpus is characterized by its use beyond a specific study or a specific research domain. The first corpus in this sense was developed in Australia for Auslan (Johnston and Schembri 2006);[3] a second collection was published online at the end of 2008 for Sign Language of the Netherlands, and efforts are under way in the UK, Germany, and Sweden.[4]

Publishing Linguistic Corpora Online

One example of a sign language corpus is that for Sign Language of the Netherlands, the *Corpus NGT* (Crasborn, Zwitserlood, and Ros 2008; Crasborn and Zwitserlood 2008). This corpus was modeled on one of the earlier Auslan corpora in terms of content and signer selection in that it aimed to include regional variation and record both more narrative and more interactive linguistic registers. In part, the same elicitation materials were used. The first release of the Corpus NGT in December 2008 contained ninety-two signers recorded in pairs, from all five regions that were identified in the lexicon projects in the 1980s and 1990s (Schermer 1990, 2003). The result was a collection of more than two thousand video clips, each containing a near-frontal view of each of two signers (figure 1).

As the introduction to this article states, it was not the project's original goal to publish the recordings for any purpose other than linguistic research. All of the video material was to be used in the online corpus of the Max Planck Institute for Psycholinguistics.[5] In this corpus, only the metadata descriptions that form part of the IMDI metadata files are publicly accessible, whereas the default state for any part of a subcorpus is "locked," that is, accessible only to the owners.[6] Additional access rights can be provided by having specific registered users listed as having access to specific files in the corpus. Thus, this online archive is an online "publication" only in the sense that users can access the data either by being an employee of the hosting insti-

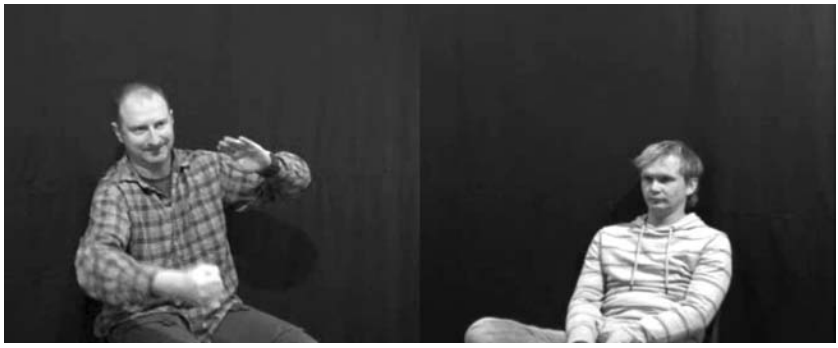

Figure 1. Appearance of the signers in the Corpus NGT publication.

tute or by subscribing to some service. It is one of the few professional ways of archiving digital video resources in linguistics while at the same time providing access over the Internet to the owners.

From the very start of the project, it became clear that various other parties might also be interested in using the data. The situation that was described earlier with regard to linguistic research—a dearth of sign language archives due to the absence of a written culture and a technological "video revolution" that has taken off only in the last decade—also holds for nonscientific areas in society. The Deaf community itself has become used to the fact that there are no signed news sources in the Netherlands other than the interpreted morning news on television, that there is no printed or videotaped sign literature, that sign language teaching and learning materials take the form of an occasional videotape or DVD, and so on. Similarly, teachers and students in the interpreting program are accustomed to the fact that there are few sign language productions outside the course materials that the school offers. Materials that include regional and age variations would aid interpreters and interpreting students in handling such variation (Crasborn and Bloem 2009; Crasborn and de Wit 2005). Parents of deaf children who want to learn NGT as quickly as possible can avail themselves of only one type of course with one type of sign language materials (similar to those used by interpreting students). For all of these groups, the various types of data in the Corpus NGT would represent a substantial addition to the existing materials, provided they be made accessible in appropriate ways. For that reason, it was decided from the start to try to publish as much of the corpus as possible online.

A more principled reason for the importance of publishing data online is that in this way research data are shared with the community in which they were collected. For decades, deaf informants or subjects for sign language research provided data for linguistic research without taking part in the research themselves or directly benefitting from any video recordings that were made. Even if they do not work as investigators in the study, deaf signers can now take part more actively by exploring the video recordings. The primary research data are available online, and it is also the goal to have a subset of the secondary research data published online, namely the annotations that contain glosses and sentence translations. A web service version of the ELAN

software, called ANNEX, can potentially serve as an access point for research data of this kind. At this time, it is unlikely that participants of the Corpus NGT will actually do so: The software is strongly oriented toward scientific use, and its interface is currently available only in English. However, the movies have been included in a Dutch web site and can be viewed by anyone.

This latter consideration, wanting to publish not only the outcomes of scientific research but also the data at their core, is one that is increasingly valued in scientific practice more generally. While the technological advances in Internet functionality and computer storage make it easier to actually share data, this change is fundamentally one in the standards and values of the scientific community at large. Similarly, the availability of the Corpus NGT and related materials as *open-access* publications fits well with current developments in scientific publishing. "Open access" refers to the "free and unrestricted online availability" (Budapest Open Access Initiative 2002) and is typically used with reference to scientific publications. The ECHO (European Cultural Heritage Online) initiative (Max Planck Society for the Advancement of Science 2003) explicitly extended this development toward scientific data, including images and video material.[7]

The publication of video recordings of deaf people signing inevitably reveals their identity to people who know them. It is not possible to hide the signers' identity by simply leaving out their names from metadata descriptions and file names. Technological manipulations of the video recordings (e.g., those that would mask a signer's face) render the material useless, given the importance of all aspects of facial expression in signing. A quick impression of such techniques is presented in figure 2, which illustrates that the videos cannot be manipulated in such a simple way before publication. A potentially more successful way of removing signers' identity from video recordings would be to convert the real person in the video to an animation character. With this in mind, a current European project (Dict-a-Sign) aims at automatic recognition and the subsequent resynthesis of signed languages by avatars. In the Corpus NGT, no effort was taken to hide the signers' identity. Instead, their explicit consent was sought to publish the video recordings "as is."

Aside from the signer's visual appearance, open-access publication

FIGURE 2. Hiding the face by image manipulation.

will also reveal the content of the interaction. Several measures were taken to ensure that the signers were aware of the nature of the publication. In the selection of participants for the recordings, it was explicitly mentioned that the goal of the project was to publish video recordings online for public access. When people arrived at the recording site, the Deaf assistant who was in charge of the recording session reminded the participants of the fact that all of the recordings would in principle be made available online. In addition, everyone was asked to be careful in what they talked about and to avoid gossip and, where possible, the mention of any names. Although language elicitation tasks were already careful to avoid personal stories, there was quite some leeway in exactly how discussions took place that might lead to privacy-sensitive remarks.

Afterward, signers were offered an opportunity to indicate that they would not want specific segments of the recordings to be published for a general audience. After the digitization and segmentation of the recordings took place, all of the participants received a DVD with their own recordings and a letter asking them to carefully look at the recordings with the privacy aspect in mind. All of the introductory segments in which the two signers introduced themselves were excluded from the open-access publication of the corpus. Some fifty sessions were excluded in response to the wishes of the signers themselves.

Finally, all of the signers signed a consent form that included the following statements; each statement could be accepted or rejected:

1. I agree to be recorded on videotape for the Corpus NGT project.
2. I agree with (parts of) the video recordings being made available through the Internet and in that way being used for research, teaching, etc. The recordings will be freely available and can be used without a charge. No money may be made out of the recordings. "Derived works" may be made: People may subtitle the movies, for example, or use images from the recordings for web sites or presentations. These derived works may not be used for making money, either.
3. I agree with these data being available forever (on the Internet or via other media).
4. I agree to being thanked by name in publications and on the web site of the project.

As several people did not accept the final statement, it was decided to not use the names of the participants anywhere. The restriction of commercial use of the published movies is enforced by a Creative Commons BY-NC-SA license, which explicitly forbids making a profit from the use of Corpus NGT materials. Each movie starts and ends with a three-second message that reminds users of this restriction. The Creative Commons licenses are becoming increasingly common in Internet publications and constitute one of the first efforts to restrict open-access publications in a way that is more general than a specific license for a specific publication or archive. For that reason, a Creative Commons license appeared to be a decent way of alerting users to the noncommercial restriction that had been promised to the signers. However, actually enforcing compliance with this license will be quite hard to do, and one has to rely on the users' fair use of the publication. In Crasborn (2008) I discuss the Creative Commons licenses and their use by the Corpus NGT in more detail.

In summary, we have aimed to make the Corpus NGT research database publicly accessible because the movies are unique in quality and quantity and should be usable not only for research but also for other purposes. We have tried to protect the signers' privacy by limiting the searchable (text) information as much as possible through the use of a restricted metadata description, and we ask for the signers' explicit consent to the online publication. In the next section I discuss

whether informed consent is really sufficient in publishing sign language recordings online.

The Value of Informed Consent in Publishing Video Data

Traditional consent forms for linguistic and psychological research have had the function of asking subjects for their permission to use data from elicitation procedures or experiments for scientific research. The concept of informed consent in science stems from medical research, where it was used to ensure patients' participation in investigations that could potentially cause harm to the participants. Informed consent referred to the doctor's obligation not to pass on information about the subject, as well as the patient's right to know what a specific investigation would consist of. From the medical sciences, the concept came to be used also in the social sciences and finally, in the 1970s, in the humanities (Fluehr-Lobban 1994). To be able to "give informed consent" to a specific investigation, a person must first have the legal right to do so. Children cannot give consent but should be represented by their parents; similarly, a legal representative should give informed consent for people who have a mental illness. Second, it should be a voluntary decision, not forced upon the persons by any means. Finally, they should be able to make an informed decision, having all of the relevant facts at their disposal (Loue 1999).

It is this last condition that poses problems when asking for informed consent with regard to video publications. The extent to which subjects are properly informed may not be as easy to judge as one would think (Schultz, Pardee, and Ensinck 1975). In the case of sign language corpora, one should ask first of all to what extent deaf people with varying levels of literacy in the researcher's spoken language (or in any spoken language for that matter) are aware of the status and impact of a consent form. Although the statements on the consent form may be explained to them in sign language (as was the case for the Corpus NGT), this does not count as a voluntary and well-informed decision if people are not fully literate and cannot comprehend the impact of a short, written document like a consent form.

Second, and more specific to the case of the publication of sign language corpora, do people understand what publication on the Internet

means? First of all, knowledge of the Internet and its developments is needed. If one has no experience browsing web pages from all over the world (but only those from one's own country), it is hard to decide whether or not video recordings of oneself should be made available to the whole world. Although this was not explicitly checked, it was the impression of our Deaf assistant that even the older signers in the Corpus NGT were acquainted with the Internet and would be able to comprehend the meaning of "worldwide availability."

Something that is much harder to evaluate for any subject is what the impact will be of the video recordings being available forever. Even if for technical, administrative, or other practical reasons the server that originally hosted a video corpus may not be functional thirty or one hundred years from now, many (segments) of the recordings will have found their way to other publications, whether online or in hard copies of DVD productions. What will it be like in the future to see recordings of oneself made twenty years earlier? Many people are familiar with the feeling of embarrassment in seeing pictures of themselves taken many years ago. It is likely that the same will happen with video recordings. Many people will take this lightly, but it is possible that others may have considerable problems with the publication of video recordings and would later like to withdraw their consent to publication. The longer materials have been online, the harder it will be to fully undo the publication of any segment that has already been published. This is inherent in the nature of the Internet, which allows for easy copying and redistributing of information, together with a license that allows users to do so. However, as was mentioned earlier, it is also due to the fact that universities have restricted legal power to enforce compliance with licenses, especially in international contexts.

Some final considerations are related to future developments that we cannot fully predict. On the one hand, it remains to be seen how sign language corpora will be used in the future and what forms possible abuse might take. This is something that we simply cannot foresee, and thus consent forms cannot cover every potential situation that may arise. Further, future technological developments may alter the way in which we look upon the publication of sign language corpora. It is not unlikely that future technologies will make it easy to recog-

nize the identity of signers on the basis of their visual appearance or even properties of their movements. Our careful efforts to refrain from mentioning the signers' names anywhere are of little value in a small group of people such as the Deaf community in the Netherlands, where people easily recognize each other, but they may also become meaningless for outside users in that way. Other seemingly unrelated technological developments could in principle impede privacy protection even further if the names of signers obtained from image or movement analysis can be linked to other databases of personal information online. As online privacy protection is under constant debate, one can assume that such developments will also cover the use of identity information in sign language corpora.

Conclusion

To summarize, there are two main considerations in evaluating the use of informed consent for the open-access publication of sign language video recordings. First, do all of the participants fully understand the value of a written consent form? This is a more general problem when dealing with subjects with various degrees of literacy. To address this situation, it is of crucial importance to reserve enough time before and after recording sessions to discuss the consent form and its importance. The second consideration relates to the fact that we are dealing with such recent technologies that the subjects may not fully understand all of the currently known consequences; moreover, upcoming technological developments may impact the online publication in a way that neither the researcher nor the subject can foresee. Here, too, it is important to explicitly mention this aspect of online publication and to double-check the "Internet literacy" of signers.

There are two reasons to go ahead with online publication even though we are dealing with some restrictions on fully informed consent. First, as regards the technological developments, we should be optimistic that Deaf people will not be singled out in these developments. New technological advances will have to be met with new restrictions and new policies of use for any Internet user, including people who understand the signing in the published videos.

Second, the publication of sign language corpora may have many

benefits for the language community itself. Those for second language learners are highlighted earlier in this article; an increase in the number of proficient L2 signers could potentially strengthen the language community. As Johnston (2004) argues for the situation in Australia, signed languages in the Western world are faced with extinction by the end of the present century due to rapid medical developments such as cochlear implants, as well as their swift adoption by the large majority of parents of deaf children. The presence of large, online video corpora will increase the visibility of signed languages and may contribute to the case for bilingual education for children with cochlear implants. Aside from these specific circumstances, for endangered spoken languages it has been argued it would be immoral to not make language materials available online, as they might contribute to the survival of the language of cultural minorities (Whalen 2001). In such contexts, the anonymity that is standard in medical research would not be applicable in cases such as storytelling, where the person who is telling the story may be as important as the content of the narrative. In the Corpus NGT, the recording of such personal or culturally sensitive narratives has been explicitly avoided or excluded from the general open-access policy.

Finally, it would be ethically sound for researchers to accept any later withdrawal of consent without hesitation and act immediately to remove the movie in question from the open-access database even if the file in case may have already been used for various purposes and even been republished in other collections.

Notes

1. This material was presented at a panel on the applied politics of deafness at Acting with Deaf People in Science, Technology, and Medicine, the joint annual meeting of the Society for the Social Studies of Science and the European Association for the Study of Science and Technology (4S/EASST), Rotterdam, the Netherlands (August 20–23, 2008).

2. The Sign Linguistics Corpora Network (SLCN) is a three-year Dutch initiative that aims to collect information on the evolution of corpus development and exploitation. In addition to a series of workshops, there will be a permanent web site, http://www.ru.nl/slcn, that gathers references and links to pertinent sources of information.

3. Http://www.auslan.org.au/about/corpus.

4. NGT: http://www.ru.nl/corpusngt; BSL: http://www.bslcorpusproject.org; DGS: http://www.sign-lang.uni-hamburg.de/dgs-korpus/homee.html.

5. Http://corpus1.mpi.nl.

6. IMDI stands for ISLE Metadata Initiative, a metadata format developed and supported by the Max Planck Institute for Psycholinguistics. See http://www.mpi.nl/IMDI. It is currently being adapted in the context of the EU project CLARIN (Common Language Resources and Technology Infrastructure); see http://www.clarin.eu for further information.

7. Within the ECHO project, I carried out a pilot study on the publication of sign language data with colleagues from the Netherlands, the United Kingdom, and Sweden (Crasborn et al. 2007). Data are published at http://www.let.ru.nl/sign-lang/echo/.

References

Bergman, B., P. Boyes-Braem, T. Hanke, and E. Pizzuto, eds. 2001. *Sign Transcription and Database Storage of Sign Information*. Special issue of *Sign Language and Linguistics* 4(1/2): 1–302.

Budapest Open Access Initiative. 2002. http://www.soros.org/openaccess/read.shtml (accessed April 20, 2009).

Crasborn, O. 2008. Open Access to Sign Language Corpora. In *Construction and Exploitation of Sign Language Corpora: Proceedings of the Third Workshop on the Representation and Processing of Sign Languages,* ed. O. Crasborn, T. Hanke, E. Efthimiou, E. Thoutenhoofd, and I. Zwitserlood, 33–38. Paris: ELRA.

———, and T. Bloem. 2009. Linguistic Variation as a Challenge for Sign Language Interpreters and Sign Language Interpreter Education in the Netherlands. In *International Perspectives on Sign Language Interpreter Education,* ed. J. Napier, 77–95. Washington, D.C.: Gallaudet University Press.

Crasborn, O., and M. de Wit. 2005. Ethical Implications of Language Standardisation for Sign Language Interpreters. In *International Perspectives on Interpreting: Selected Proceedings from the Supporting Deaf People Online Conferences 2001–2005,* ed. J. Mole, 141–50. Brassington, UK: Direct Learn Services.

Crasborn, O., T. Hanke, E. Efthimiou, E. Thoutenhoofd, and I. Zwitserlood, eds. 2008. *Construction and Exploitation of Sign Language Corpora: Proceedings of the Third Workshop on the Representation and Processing of Sign Languages*. Paris: ELRA. http://www.lrec-conf.org/proceedings/lrec2008/workshops/W25_Proceedings.pdf.

Crasborn, O., J. Mesch, D. Waters, A. Nonhebel, E. v. d. Kooij, B. Woll, and B. Bergman. 2007. Sharing Sign Language Data Online: Experiences from the ECHO Project. *International Journal of Corpus Linguistics* 12(4): 535–62.

Crasborn, O., and I. Zwitserlood. 2008. The *Corpus NGT:* An Online Corpus for Professionals and Laymen. In *Construction and Exploitation of Sign Language Corpora: Proceedings of the Third Workshop on the Representation and Processing of Sign Languages,* ed. O. Crasborn, T. Hanke, E. Efthimiou, E. Thoutenhoofd, and I. Zwitserlood, 44–49. Paris: ELRA.

———, and J. Ros. 2008. *The Corpus NGT: An Open Access Online Corpus of Sign Language of the Netherlands.* http://www.ru.nl/corpusngt.

Fluehr-Lobban, C. 1994. Informed Consent in Anthropological Research: We Are Not Exempt. *Human Organization* 53(1): 1–10.

Johnston, T. 1991. Transcription and Glossing of Sign Language Texts: Examples from Auslan (Australian Sign Language). *International Journal of Sign Linguistics* 2(1): 3–28.

———. 2004. W(h)ither the Deaf Community? Population, Genetics, and the Future of Australian Sign Language. *American Annals of the Deaf* 148: 358–75.

———, and A. Schembri. 2006. Issues in the Creation of a Digital Archive of a Signed Language. In *Sustainable Data from Digital Fieldwork,* ed. L. Barwick and N. Thieberger, 7–16. Sydney: University of Sydney Press.

Loue, S. 1999. *Textbook of Research Ethics: Theory and Practice.* New York: Kluwer Academic Publishers.

Max Planck Society for the Advancement of Science. 2003. European Cultural Heritage Online (ECHO): Open Access Infrastructure for a Future Web of Culture and Science. http://echo.mpiwg-berlin.mpg.de/home (accessed April 21, 2009).

Schermer, T. 1990. In Search of a Language: Influences from Spoken Dutch on Sign Language of the Netherlands. PhD diss., University of Amsterdam.

———. 2003. From Variant to Standard: An Overview of the Standardization Process of the Lexicon of Sign Language of the Netherlands (SLN) over Two Decades. *Sign Language Studies* 3(4): 96–113.

Schultz, A., G. Pardee, and J. Ensinck. 1975. Are Research Subjects Really Informed? *Western Journal of Medicine* 123: 76–80.

Whalen, D. 2001. How Can We Ethically Put Language on the Web? *Endangered Languages Fund* (newsletter) 5(1): 1–4.