Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages, pages 95–100
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Approaches to the Anonymisation of Sign Language Corpora

Amy Isard
Department of Languages, Literature and Media
University of Hamburg
amy.isard@uni-hamburg.de

Abstract

In this paper we survey the state of the art for the anonymisation of sign language corpora. We begin by exploring the motivations behind anonymisation and the close connection with the issue of ethics and informed consent for corpus participants. We detail how the the names which should be anonymised can be identified. We then describe the processes which can be used to anonymise both the video and the annotations belonging to a corpus, and the variety of ways in which these can be carried out. We provide examples for all of these processes from three sign language corpora in which anonymisation of the data has been performed.

Keywords: sign language corpora, corpus anonymisation

1. Introduction

The purpose of anonymisation is to ensure that no personal information is shared for which the person concerned has not given their informed consent. The discussion of what exactly informed consent is, and how to obtain it, is not a simple one (Crasborn, 2010; Rock, 2001; McEnery and Hardie, 2011; Singleton et al., 2014; Schembri et al., 2013). The issues vary depending on among other things the size of the community in which the corpus is collected, the nature of the corpus content and the technological background of the subjects, and it is important to consult the subjects about what they would find appropriate. When describing data collection with the shared signing community in Adamorobe, Kusters (2012, p. 32) observes: “As for anonymity it appeared that people were happy for me to use their real names. The idea of changing their names in ‘a book that is about them’, seemed very odd to them.” Singleton et al. (2014, supplementary material) asked Deaf focus group participants for suggestions about how to use material in research presentations while maintaining anonymity, and they suggested the use of avatars or actors to reproduce the data, or digital editing which could obscure the subject’s identity.

Conversations in sign language corpora also often contain mentions of third parties, who are known to the corpus participants but have not been asked for or given any kind of consent for information about them to be shared publicly. Particularly when small communities are involved, it is often easy to identify a person from minimal amounts of information, and care should therefore be taken to obscure as much of this information as possible if videos and annotations are going to be available to the public. Before any analysis or annotation work is carried out on a corpus, participants should always be given a copy of their own recordings and allowed the further opportunity to refuse consent for all or any parts of the recordings to be shown or used in any way.

The process of anonymisation is expensive and time-consuming, and many corpus projects have taken the decision to publicly release only parts of the data where no personal information is revealed, or to ensure that informed consent has been acquired to the best standard possible, and/or that anyone who has access to the data has signed a confidentiality agreement and understands exactly how the data may be used for further research.

In this paper, we describe what the options are once the decision to carry out anonymisation has been taken, and various ways in which these can be implemented. Throughout the rest of the paper, examples of the anonymisation processes and techniques used by the three corpora briefly described below will be used.

The DGS Corpus is a corpus of German Sign Language (DGS). It consists of 560 hours of video dialogues, and about 50 hours has been made available as the Public DGS Corpus[1] (Jahn et al., 2018). The data was elicited using 18 different tasks, some of which involved free conversation where personal information about third parties was sometimes mentioned. The Public DGS Corpus video and annotations have been anonymised to remove references to which would allow the identification of third parties.

The NGT Corpus is a corpus of Dutch Sign Language (NGT). It consists of dialogues between 92 participants and is available online[2] (Crasborn and Zwitserlood, 2008). A number of different elicitation tasks were used and some of the conversations involve references which could identify third parties. The available annotations have been anonymised but the video has not.

The Rudge Corpus is a small corpus of British Sign Language (BSL) collected by Luke Rudge for his PhD thesis on the topic of the use of Systemic Functional Grammar in the analysis of BSL (Rudge, 2018). There were 12 participants who gave pre-prepared presentations about a prominent period in their lives, which sometimes revealed personal information. The videos and annotations have been anonymised but they are not publicly available.

2. What to Anonymise

In sign language corpora, it is impossible to completely anonymise the video data, because both the face and hands

[1] `http://ling.meine-dgs.de`

[2] `https://www.ru.nl/corpusngten/`

of the participants must be fully visible for the content to be understandable (Quer and Steinbach, 2019; Hanke, 2016; Crasborn, 2010). Chen Pichler et al. (2016, page 32) note that: “there appears to be virtually unanimous agreement that total anonymization, long taken as a standard practice for medical data, is not feasible for language data that include audio and/or video components”.

Although it is not possible to completely conceal the identities of the participants in a sign language corpus, it is nonetheless necessary to ensure that as few of their personal details are revealed as possible. In addition, care must be taken to obscure personal information of third parties who are mentioned during the dialogue, if it could lead to their identification. These third parties will not have had the opportunity to give their informed consent for any sort of appearance in the corpus.

There are two main situations in which the anonymisation of sign language corpora is carried out:

- anonymisation of a whole corpus for wider distribution to a larger team or outside researchers
- anonymisation of single words or phrases for use in settings such as a conference talk, seminar or sign language dictionary

In both cases, it is first necessary to identify which information needs to be anonymised. In a small corpus it may be possible to make the selection by watching all the videos, but in a larger corpus it maybe helpful to use some automatic processing. The anonymisation of videos is described in Section 3, and of annotations in Section 4.

3. Anonymisation of Video

There are a number of different ways in which video can be anonymised. These can be divided into two categories, those which *conceal* all or part of a video, and those which *reproduce* a video. Concealing can be effected on part or all of a video frame with the use of blurring or pixellation, or by obscuring the image entirely. Reproduction can be carried out by using either an actor or a computer-generated avatar.

These two approaches are generally used for different purposes. Reproduction can conceal the identity of the signers themselves, while concealing preserves the anonymity of third parties by hiding references to people or places. No detailed studies have been published about the extent to which reproduction affects the viewer’s understanding of a sign language video, or what level of blurring is necessary to ensure that the movements cannot be distinguished. In the related area of spoken dialogue research, the CASE corpus of Skype dialogues experimented with video anonymisation using *Adobe Premiere* pixel, art, and transformation filters, and chose a contour filter. In control tests, they discovered that when this filter was used, subjects did not recognise themselves (Diemer et al., 2016).

3.1. Concealment

Concealment can be used on just part of the image of a video, and usually over a small time frame. The viewer’s experience is not hugely disrupted, as only a sign or two will be concealed. Inevitably some information will be lost, but this can be kept to a minimum. The concealment can be carried out by blackening all or part of the image ()adding one or more black rectangles), or by blurring or pixellating all or part of the image to such an extent that the signing or mouthing is no longer recognisable.

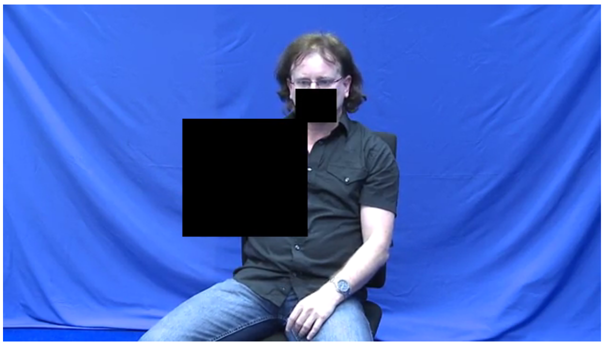

Figure 1: Screenshot from the DGS Corpus, anonymised through blackening with one black rectangle over the mouth and cheeks and another over the right hand and arm and the top right portion of the torso.

3.1.1. Blackening

In the Public DGS Corpus mentions of sensitive information in videos are anonymised by blackening sections of the image (Bleicken et al., 2016). The timings from the annotation tiers (see Section 4) are used to identify the relevant timespan. Experiments were carried out which showed that if the whole timespan was blackened, this invalidated a whole sentence for linguistic analysis, because it disturbed suprasegmental signals. They therefore imposed one or more black rectangles on the image, to cover the mouth, one or both hands and/or the trunk, depending on the position of the sign. Experiments also showed that blackening was less disturbing to viewers than pixellation. OpenPose analysis (Cao et al., 2017) had already been carried out on the corpus (Schulder, 2019), providing machine-readable information on the location of various body parts, such as hands, shoulders, and mouth, so this was used to find the location of the relevant body parts, and the size and shape of the rectangles were then adjusted by hand. An example screenshot is shown in 3.1.1, where the mouth, cheeks, right hand and right arm of the signer have been hidden, along with a portion of the torso in front of which the sign was being performed.

3.1.2. Pixelation

For the Rudge corpus, the author went through the video recordings and noted where participants had signed a proper name of a person, specific location or any other information which could identify a third party. The video was then loaded into editing software such as *Final Cut Pro* or *Adobe After Effects*, and a local blur or pixellation filter was applied to the signer’s hands and mouth for the duration of the relevant sign, which was normally only a few tenths of a second during fluent signing (Rudge, personal communication, January 2020). This ensured that any third party information had been removed before the recordings were passed to other researchers for annotation. No screenshots are available as participants did not give consent for any images to be shown to people outside the initial small research group.

3.2. Reproduction

Reproduction of a corpus can in theory be carried out by either humans or computer-generated avatars. Some corpus examples where human actors have been used are described in Section 3.2.1 and the steps which would be necessary for avatar reproduction in Section 3.2.2.

3.2.1. Actors

For total anonymity, short examples from a corpus can be reproduced by a human actor. In this case complete anonymity is assured, but there are several disadvantages as a result. The process is very labour-intensive, requiring not only the time of the signer but also of a studio and technicians to carry out the recording. In addition, no matter how well the second signer copies the original, some information will be lost. Performativity is a vital part of sign language and it is impossible to fully separate the affective and grammatical functions of facial expressions.

The participants in the Rudge corpus had agreed only to their recordings being seen by the author and a limited number of other researchers who worked on verification of the data. Because the thesis is publicly available, examples used in it were reproduced by the author or another signer, to preserve the anonymity of the original participants (Rudge, 2018 and personal communication, January 2020).

The DGS Corpus is being used in the compilation of a Dictionary of German Sign Language and the preference is to use examples taken directly from the corpus, for the reasons discussed in detail in Langer et al. (2018). However, in very occasional cases where the dictionary compilers want to use an example which contains personal information about a third party, they re-record the example with a signing model and replace any personal names in the re-recording and the associated translation with a common German family name.

3.2.2. Avatars

In practice, although avatars have been improving rapidly in quality, no large-scale avatar reproduction has been carried out. A survey of the state of the art in sign language avatars can be found in (Bragg et al., 2019). The are a number of technical problems with the use of avatars for sign language, and some of these are related to the process of creating the content and ensuring that the correct manual and non-manual gestures are created. In the case of reproduction these particular issues are avoided, because the data for the avatar comes directly from the original videos. The problems of designing avatars which are acceptable to the Deaf community in terms of appearance and comprehensibility remain, and it is essential that the acceptability of avatars be systematically reviewed and assessed before they are used (Kipp et al., 2011).

In order to use avatars for reproduction, the original videos must first be processed using pose estimation software, which can identify particular body parts including hands, arms, and facial features. A visual representation of an OpenPose analysis from the DGS corpus (Schulder, 2019) is shown in Figure 3.2.2, illustrating the keypoints identified by the software and lines between the points to indicate logical connections between them. However, OpenPose only produces two-dimensional images, and additional (extremely time and resource intensive) processing is required to reconstruct three-dimensional images (Xiang et al., 2019). The resulting machine-readable information on the location of various body parts could then be used to animate an avatar which would reproduce the desired data, but as far as we are aware, no sign language avatar has so far been tested on this output.

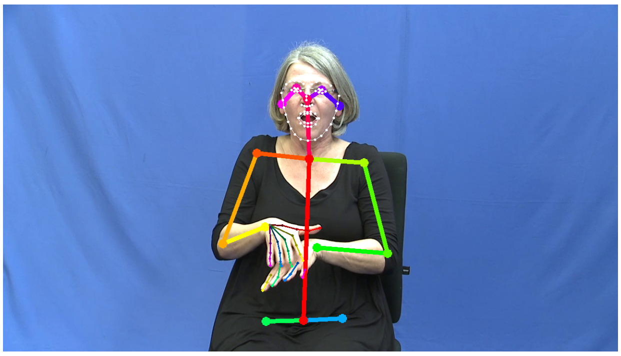

Figure 2: Visual representation of the pose information provided by OpenPose, computed for a video from the DGS-Korpus project. Sets of keypoints are generated for the body, the face and each hand. Lines between the points are added to the visual representation to indicate the logical connection between individual keypoints.

3.2.3. OpenPose data

If OpenPose data are made publicly available, they must also be anonymised, to the same level as the videos on which they were based. If the data were later used, for example to animate an avatar, they could make personal information visible. OpenPose data are available to download as part of the Public DGS Corpus (Schulder, 2019), and they have been anonymised to remove keypoints for timespans which were previously chosen for anonymisation as described in Section 4.1. It is possible to differentiate between keypoints which have been anonymised and those which are missing because the body part is temporarily hidden (for example when a person puts a hand behind their head), so that if the OpenPose data were used to animate an avatar, anonymised keypoints could be covered by a black square, as with video blackening (see Section 3.1.1).

4. Anonymisation of Annotations

Before the anonymisation of annotations can be carried out, the sensitive names must first be identified. In a small corpus, this may have been done by watching the video data, but where many hours of video have been translated and annotated by a team of researchers, automatic methods can also be used.

4.1. Name Identification Methods

For the Rudge corpus, names were found by manual inspection of the videos (see Section 3).

In the NGT corpus, information which had been manually annotated in the gloss and mouth tiers was used to identify names which needed to be anonymised (Crasborn and Bank, 2015).

The DGS-Korpus project tested a subset of the DGS corpus to see how reliable different techniques were for finding sensitive items which should be anonymised (Bleicken et al., 2016). Because German translations had already been carried out, they could use computational linguistic tools for German which are available through Weblicht (Hinrichs et al., 2010) as pre-defined chains.

They used four approaches and compared the results for each to a ground truth defined as the sum of the names correctly identified by each technique. The four approaches which they used were:

- Manual inspection of the videos by a deaf annotator who was asked to mark every occurrence of a name
- Extraction of potential names from the annotations, which were then checked against the German translations; when a match was found, a manual inspection was carried out
- Use of named entity recognition on the German translations
- Checking mouthing annotations and translations against name lists

When comparing the final outcomes of all the methods, they found that the most effective process was to combine the automatic methods with a one-pass manual inspection.

The DGS-Korpus project found that they were more conservative in their selection of data which needed to be anonymised than the participants themselves had been after reviewing their own recordings. They decided therefore that it was unfair to make the participants entirely responsible for these decisions, and better to be more cautious, and carry out more anonymisation rather than less, in an effort to prevent any identifiable information on third parties being released accidentally.

Once the names to be anonymised have been identified, the actual anonymisation can be done using either *categorisation* or *pseudonymisation*. Pseudonymisation involves the use of replacement names (Section 4.2). In categorisation, a name is usually replaced by a string indicating the type of proper name plus a numeric identifier, so that subsequent mentions in the same dialogue can be seen to be referring to the same entity (Section 4.3).

4.2. Pseudonymisation

When pseudonymisation is carried out, the pseudonyms can be chosen to match the original names on as many levels as desired. This could for example involve choosing replacement cities of approximately the same size, or family names which originate from the same geographical region. Anonymisation with pseudonyms was for example carried out in the spoken German FOLK corpus (Schmidt, 2016; Winterscheid, 2015). One disadvantage of this approach is that it can be very time consuming as time must be spent choosing replacement names and making sure that they fit all of the chosen criteria. There are currently no sign language corpora for which a description of anonymisation using pseudonyms is available. Issues to consider would include the question of how to define “similar” names in terms of sign language phonology.

4.3. Categorisation

Categorisation is a quicker and simpler process than pseudonymisation because it is only necessary to identify the type of a proper name in order to create its replacement.

In the NGT corpus, glosses and annotation tiers are anonymised so that it will not be possible for anyone to make a simple automatic search for names. All glosses which refer to participants and other people who are not considered to be in the public domain are replaced by the type `*NAMESIGN`. In mouthing and translation tiers, they are replaced by the type `*eigennaam` (Crasborn and Bank, 2015).

In the Rudge corpus, the timestamps from the manual analysis of the video data (see Section 3.1.2) were used to find places in the translation and annotations where names, locations and other personal data needed to be anonymised. They were replaced with types such as `[NAME]` or `[LOCATION]`. If there were multiple instances of anonymisation in the same clause or in quick succession, a suffix was added of the form `[NAME-a]`, `[NAME-b]`, etc. so that any following indicating verbs or signs requiring more complex spatio-kinetic features (e.g., placement in the signing space) could still be understood in spite of the visual noise (Rudge, 2018 and personal communication, January 2020).

The DGS-Korpus project examined each person name to determine whether it belonged to someone for whom information is already available in the public domain, such as television personalities or politicians, whose names would not then be anonymised. They also defined a population threshold above which places were considered to be large enough to not require anonymisation. Proper names in the translation and mouthing annotations, and most of the gloss tier, were replaced by numbered placeholders of the form `Name#1`, `Name#2`, etc. so that it is still possible to tell when the same person or place is referred to more than once.

5. Final Thoughts

It must always be kept in mind that in a large corpus it is basically impossible to ensure that all possible identifiable information has been removed, and that this must be made clear to the participants as part of the process of obtaining informed consent. For example, in one dialogue from the Public DGS corpus (English translation shown below), a place name is anonymised, but two sentences later it is mentioned that it is the previous residence of a princess from the 18th century who, as a person in the public eye, would not normally have her name anonymised:

> My hometown Place#1 also has a small tourist attraction.
> There used to be a castle right where the German Catholic Church is located today.
> The Austrian princess Elisabeth used to live there.

It would therefore be theoretically possible for someone who comes from the same area or has a thorough knowledge of the history of the region to figure out the name of the participant's home town. To avoid this, the name of the princess would then also have to be anonymised, and possibly even her nationality, but at some point a decision has to be made about how far to continue the process, and in this case, it was decided that the name of the princess would not be anonymised.

6. Acknowledgements

This work was supported by the BMBF (German Federal Ministry of Education and Research) Project QUEST: Quality-Established[3].

7. Bibliographical References

Bleicken, J., Hanke, T., Salden, U., and Wagner, S. (2016). Using a Language Technology Infrastructure for German in order to Anonymize German Sign Language Corpus Data. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3303–3306, Portorož, Slovenia.

Bragg, D., Koller, O., Bellard, M., Berke, L., Boudreault, P., Braffort, A., Caselli, N., Huenerfauth, M., Kacorri, H., Verhoef, T., Vogler, C., and Ringel Morris, M. (2019). Sign Language Recognition, Generation, and Translation: An Interdisciplinary Perspective. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '19, pages 16–31, Pittsburgh, PA, USA.

Cao, Z., Simon, T., Wei, S.-E., and Sheikh, Y. (2017). Realtime multi-person 2D pose estimation using part affinity fields. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7291–7299, Honolulu, HI, USA.

Chen Pichler, Deborah, D., Hochgesang, J., Simons, Doreen, D., and Lillo-Martin, Diane, D. (2016). Community Input on Re-consenting for Data Sharing. In *Proceedings of the Seventh Workshop on the Representation and Processing of Sign Languages: Corpus Processing at LREC 2016*, Portorož, Slovenia.

Crasborn, O. and Bank, R. (2015). Corpus NGT Anonymisation Protocol. `https://www.academia.edu/40438732/Corpus_NGT_Anonymisation_Protocol`.

Crasborn, O. A. and Zwitserlood, I. E. P. (2008). The Corpus NGT: An Online Corpus for Professionals and Laymen. In *Proceedings of the Third Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora at LREC 2008*, pages 44–49, Marrakech, Morocco.

Crasborn, O. (2010). What Does "Informed Consent" Mean in the Internet Age? Publishing Sign Language Corpora as Open Content. *Sign Language Studies*, 10(2):276–290.

Diemer, S., Brunner, M.-L., and Schmidt, S. (2016). Compiling computer-mediated spoken language corpora: Key issues and recommendations. *International Journal of Corpus Linguistics*, 21(3):348–371.

Hanke, T. (2016). Towards a Visual Sign Language Corpus Linguistics. In *Proceedings of the Seventh Workshop on the Representation and Processing of Sign Languages: Corpus Mining at LREC 2016*, pages 89–92, Portorož, Slovenia.

Hinrichs, E., Hinrichs, M., and Zastrow, T. (2010). WebLicht: Web-Based LRT Services for German. In *Proceedings of the ACL 2010 System Demonstrations*, pages 25–29, Uppsala, Sweden.

Jahn, E., Konrad, R., Langer, G., Wagner, S., and Hanke, T. (2018). Publishing DGS corpus data: Different Formats for Different Needs. In *Proceedings of the Eighth Workshop on the Representation and Processing of Sign Languages: Involving the Language Community at LREC 2018*, pages 83–90, Miyazaki, Japan.

Kipp, M., Heloir, A., and Nguyen, Q. (2011). Sign Language Avatars: Animation and Comprehensibility. In Hannes Högni Vilhjálmsson, et al., editors, *Intelligent Virtual Agents*, Lecture Notes in Computer Science, pages 113–126, Berlin, Heidelberg. Springer.

Kusters, A. (2012). Being a deaf white anthropologist in Adamorobe: Some ethical and methodological issues. In *Sign Languages in Village Communities: Anthropological and Linguistic Insights*, pages 27–52. De Gruyter Mouton, Berlin, Boston.

Langer, G., Müller, A., Wähl, S., and Bleicken, J. (2018). Authentic Examples in a Corpus-Based Sign Language Dictionary – Why and How. In *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts.*, pages 483–497, Ljubljana, Slovenia.

McEnery, T. and Hardie, A. (2011). *Corpus Linguistics: Method, Theory and Practice*. Cambridge Textbooks in Linguistics. Cambridge University Press.

Quer, J. and Steinbach, M. (2019). Handling Sign Language Data: The Impact of Modality. *Frontiers in Psychology*, 10.

Rock, F. (2001). Policy and Practice in the Anonymisation of Linguistic Data. *International Journal of Corpus Linguistics*, 6(1):1–26.

Rudge, L. A. (2018). *Analysing British Sign Language through the Lens of Systemic Functional Linguistics*. Ph.D. thesis, University of the West of England. `https://uwe-repository.worktribe.com/output/863200`.

Schembri, A., Fenlon, J., Rentelis, R., Reynolds, S., and Cormier, K. (2013). Building the British Sign Language Corpus. *Language Documentation & Conservation*, 7:136–154.

Schmidt, T. (2016). Construction and Dissemination of a Corpus of Spoken Interaction - Tools and Workflows

[3] `https://www.slm.uni-hamburg.de/en/ifuu/forschung/forschungsprojekte/quest.html`

in the FOLK project. *Journal for Language Technology and Computational Linguistics*, 31(1):127–154.

Schulder, M. (2019). OpenPose in the Public DGS Corpus. Project Note AP06-2019-01, Institute for German Sign Language, Hamburg University, Hamburg, Germany. `https://www.sign-lang.uni-hamburg.de/dgs-korpus/arbeitspapiere/AP06-2019-01.html`.

Singleton, J. L., Jones, G., and Hanumantha, S. (2014). Toward Ethical Research Practice With Deaf Participants. *Journal of Empirical Research on Human Research Ethics*, 9(3):59–66.

Winterscheid, J. (2015). Maskierung. Working Paper, Institut für Deutsche Sprache, Mannheim. `https://ids-pub.bsz-bw.de/frontdoor/deliver/index/docId/3904/file/Winterscheid_Maskierung_2015.pdf`.

Xiang, D., Joo, H., and Sheikh, Y. (2019). Monocular total capture: Posing face, body, and hands in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10965–10974, Long Beach, CA, USA.