Using a Language Technology Infrastructure for German in order to Anonymize German Sign Language Corpus Data

Julian Bleicken, Thomas Hanke, Uta Salden, Sven Wagner
Institute of German Sign Language and Communication of the Deaf, University of Hamburg
Binderstraße 34, 20146 Hamburg, Germany
E-mail: {julian.bleicken,thomas.hanke,uta.salden,sven.wagner}@sign-lang.uni-hamburg.de

Abstract

For publishing sign language corpus data on the web, anonymization is crucial even if it is impossible to hide the visual appearance of the signers: In a small community, even vague references to third persons may be enough to identify those persons. In the case of the DGS Korpus (German Sign Language corpus) project, we want to publish data as a contribution to the cultural heritage of the sign language community while annotation of the data is still ongoing. This poses the question how well anonymization can be achieved given that no full linguistic analysis of the data is available. Basically, we combine analysis of all data that we have, including named entity recognition on translations into German. For this, we use the WebLicht language technology infrastructure. We report on the reliability of these methods in this special context and also illustrate how the anonymization of the video data is technically achieved in order to minimally disturb the viewer.

Keywords: sign language, corpus, named entity recognition, anonymization

1. Introduction

In its first phase, the DGS Korpus project collected a corpus of conversations in DGS (German Sign Language) from 165 pairs of informants totaling in multi-camera video recordings of 825 hours (cf. Nishio et al., 2008). In the second phase, basic annotation as well as translations were added. As it is the aim of the project to provide language data for linguistic research and at the same time to contribute to the cultural heritage of the sign language community, parts of the corpus shall be made available to the public via a website. This of course raises the question as to what part of the data is to be anonymized before publication and how exactly this is to be accomplished.

Sign language users not only have their hands as articulators, but other parts of the body are information channels as well, such as eyebrows, eye gaze, mouth, head movement etc. Hiding the face in order to make identification of the signer impossible therefore is not an option. The state of the art in avatar technology does not allow to faithfully reproduce video-recorded sign language data with economical limits to work invested. So there is no alternative to making the signers in the DGS corpus data fully visible and recognizable to the viewer. Fortunately, most participants are very proud of their involvement in the project and have agreed to their videos being published on the web.

As our informants typically spent around seven hours on-site, being occupied with a variety of diverting tasks, many were fully engaged in the conversations and completely forgot about the surroundings, i.e. that there were filmed. This resulted in close-to-natural conversation, often revealing details about themselves or other persons not really suitable to be made public. (We consider the fact that this happened rather regularly as a success of our data collection process.) In order to identify passages not suitable for publication, as a first step, we made the recordings available for the informants by sending them DVDs that they could review at their home's TV set and note down timespans they did not want to see published.

From the beginning on, we asked our annotators to pay attention to content not suitable for publication due to inappropriate language or contents, and it turned out that we wanted to exclude more data than what the informants had asked. So the part of the corpus to be published avoids such content.

However, the remaining parts still contains lots of references to third persons. One typical example is once the two informants had identified some overlap in their social nets, they started talking about persons they both knew. In this case, good practice (cf. Rock 2001) requires us to remove elements from the presentation that allow identification of the person being talked about, such as the name, but also some geolocations, as small places mentioned in the context of that person might be hints as we are dealing with a small community where mentioning the living place of a person might be all you need to identify that person.

2. Names in DGS

Names as used in DGS can be as easy to identify as fingerspelling the person's or location's name (i.e. by spelling the name in its Latin alphabet form in the air, more or less letter by letter), it might be name signs not related to the German-language name (often showing physical properties of the named) or related to the German (often the case with family names meaning some profession in German). More subtle first references to a person found in the corpus are indexical signs pointing to a non-present person arbitrarily located in signing space, articulated with the mouthing of the German name.

If we already had reliable detailed annotations of all parts to be published, one could hope that all third-person or location references could be identified from the annotation. This is not the case, however, with the basic annotation achieved so far.

From the examples given above what might constitute a person reference in the sign language data, it becomes clear how difficult it is to tag person references by just attentively watching the video.

3. Identification of Named Entities

Having the translations into German available as well, we decided to additionally use these for name identification. This resulted in four different approaches to be compared in the following:

- Extracting name reference candidates from the annotation and manually inspecting these (this covers fingerspelling as well as ordinary signs meaning concepts that are often used as names in Germany (including professions, plants and some others),
- Having a person watch the video and tag name references,
- Using named entity recognition on the (time-aligned) translations into German.
- Checking mouthing annotations as well as translations against name lists with first names, last names and German geolocations.

We provide data on an experiment with a part of the corpus detailing which percentage of the "ground truth" names are detected with each method. Lacking any better method, the ground truth is constructed as the sum of all correct name hits contributed by the four different approaches. For the evaluation of this experiment, extensive additional checking of the data revealed no deficits of the so constructed ground truth.

It is obvious that any method working on the translation of the language data instead of on the original data will be skewed. Translations errors and different strategies for referencing between German and DGS play a role. On the one hand, this resulted in name references in the video that could not be found via the translations, because they were either replaced by a pronoun in the translation (two cases in our sample), forgotten (one case) or wrongly translated (one case). These cases were counted as false negatives for the translation-based approaches. On the other hand, an indexical sign that implicitly referred to a person or a location was explicitly translated by mentioning the name reference (one case in our sample). Since the name did not appear in the video, it could not be detected by the visual inspection and therefore counted as false negative for this approach.

We examined 31 minutes of the corpus data in total from three different conversations. As we wanted different DGS dialects to be covered in the sample, one pair of informants was from the very north of Germany, while the other two pairs of informants were from the south of Germany. The dialects are reflected not only in varying signs, but also in divergent mouthings. This might make the identification of names harder for our staff member, who watched the videos, and came from the north of Germany, because he was not used to mouthings from the south.

3.1 Annotation-based Inspection

Fingerspelling and signs exclusively used as name signs are easy to spot as they use special glossing conventions. Further candidates to test are names that have been marked in the lexical database as usable for name signs or as conventionalized name signs for cities or persons well known in the Deaf community. From these entries in the database we generated a list that contained the names in full length. Additionally, multi-part names, e.g. first name and surname, were split and each part of the name was inserted into the list separately. This list was checked against the German translations of the videos, matches were considered as name reference candidates and visually inspected.

This list of concepts provided 53% true positives, 5% false positives and 47% false negatives. It did not provide additional name reference candidates to the other approaches. However, a match with the list of concepts might help to decide whether or not a name should be anonymized because the list contains mainly well-known persons or places.

3.2 Manual Inspection

A deaf annotator was asked to view the video and to mark each occurrence of a name. From the examples given above, it is clear that a good understanding of the signed content is crucial. The annotator had not seen the experiment data before and was allowed to stop and review the video as often as necessary. For the inspection of the 31 minutes of video included in the sample at hand, the annotator spent 2.5 hours, a time long enough to include fatigue effects on the results. Nineteen minutes of the sample were signed in an unfamiliar dialect for the annotator. The manual inspection revealed 93% true positives, 5% false positives and 7% false negatives. In total there were only four false negatives, one to be neglected, because a name was mentioned in the translation only (see above). However, one of these cases was a name that had to be anonymized. Assuming we had relied only on the manual inspection, we had missed this entity. As expected, it was harder for the annotator to detect names when informants signed in an unfamiliar dialect. But even if the concrete meaning of a name was not understood, the entity was still identified as a name.

It is striking that the annotator marked several institutional names as name references. These were included in the "ground truth" if the name of the institution was correct and complete. This was true for only one name in the sample that was exclusively detected by the manual inspection, but should have been detected by named entity recognition approaches as well. One has to admit, however, that most of the institutional names were specific to the Deaf community and the NER approaches might not be trained that way. Complete organization or institutional names usually contain a location information, e.g. '*Deutscher* Gehörlosenbund' (*German* Association of the Deaf), 'Gehörlosenverein *Hamburg*' (Deaf Club *Hamburg*). Through these location information organization names should easily be detected by NER. However, in conversational language the location information is often dropped inhibiting the detection. In our sample one name of a small city had to be anonymized. The informant reports that the chairman of the Deaf club in Smallcity had a bad reputation. Because both function and city name are given, that person's identity could easily be revealed. Therefore it is necessary to consider even shortened institutional names as candidates for anonymization. Otherwise one would miss passages as described in case the informant had reported the same information in two sentences, e.g. 'I live in Smallcity. The chairman of the Deaf club there has a bad reputation.'. The decision of the annotator to include incomplete institutional names into the candidates list was therefore useful.

Finally, the annotator marked a number of event descriptions that seem to be rather general as names. Apparently, for him the event descriptors, such as 'the Cologne open day', unambiguously identified specific events so they could be considered names. Although this is again a case not influencing the results of our test, it is relevant for the general task of anonymization as the event might be the anchor for person references.

3.3 Named Entity Recognition on Translations

For named entity recognition, we implemented calling pre-defined WebLicht (Hinrichs et al., 2010) chains into our annotation environment iLex (Hanke/Storz 2008) and ran our data through two different named entity recognizers available in WebLicht. As we were well aware that most such systems are trained on written text, while we feed them with translations of face-to-face communication, we had to expect some errors, mostly false negatives. On our sample the WebLicht named entity recognizers produced 86% true positives, 26% false positives and 14% false negatives. As mentioned above, four of the 8 false negatives can be neglected, as they derive from the fact that original language data and the translations the named entity recognition ran through were not identical. Nevertheless, the remaining false negatives are still too much if one wanted to use this approach alone.

3.4 List of Names

The list of names comprised of the 2700 most common last names in Germany, first names that can be given to children in Germany, as well as some 165000 geolocation names (from geonames.org). The geolocations were further manipulated: Multi-part expressions were split as described for the concept list in 3.1 and additionally extended by plural and genitive endings. Checking against the name list results in many false positives, especially with names identical to rather common German words, like the name of a small German river, Sie, identical in writing to a pronoun. Some of these names that produced too much false positives were removed from the list in order to facilitate follow-up work. Checking against the name list produced 70% true positives, 258% false positives and 30% false negatives. The false negatives contained mainly foreign names or names written in a rarely used way, as well as bigger geolocations like continents. At least in the sample at hand the name list did not contribute any name finding that was not also found by another approach. In order to improve the output of the list further, names should be removed that generate a lot of false positives. Additionally, institutional names often used in conversational DGS (see 3.2) could be added to the list.

3.5 Results

Not surprisingly, manual inspection by competent signers yields the best results of the methods investigated, but is also (with an effort of five times real time) rather costly. It has to be noted, however, that automatic procedures with high rates of false positives cause substantial costs for manually identifying the false alarms as such.

At least in this experiment, the additional manual checking did not find extra cases, so that our preliminary conclusion is that a combination of a single one-pass manual inspection with the other methods discussed is good enough.

The combination of methods not only achieves slightly better results for the original language data than manual inspection alone, but also provides a good chance to catch names in the translation not present in the original without spending another manual inspection on the translation.

4. Applying Anonymization to the Data

Once named entity references are identified, it needs to be decided whether they need to be removed for the data to be published. This is not the case if the reference is to persons of public interest, whether for the community at large or the Deaf community in particular. Here, we followed Sharoff (2006) by assuming that participants have no personal relation to politicians etc. For members of the Deaf community, this would not be a valid assumption. So here we check every case whether information provided about the third person is in the public anyway or if the information stems from private contacts. The same procedure was applied to small organizations. For references to places, we defined a population size threshold above which we considered these uncritical. For smaller places, we manually checked whether the place reference could contribute to re-identification of any third person mentioned.

Now the question remains how to remove the data from the corpus. In the case of translations and mouthing annotations, named entities are replaced by numbered placeholders, e.g. Name#1 in order for the user to be able to follow co-references. In most cases, the same applies to the gloss tier.

For the video, the annotation determines the timespan to be manipulated. However, some experiments showed that completely blackening that timespan invalidates the whole sentence for further linguistic analysis as suprasegmental signals are disturbed. Therefore, we defined several options how to manipulate a stretch of video sufficient to make the sign or mouthing component unrecognizable:

- In the case of mouthing, only the mouth including cheeks and the chin is to be hidden.
- In the case of fingerspelling, only the dominant hand and the surrounding covering the sideways and downwards movements potentially occurring need to be covered.
- For signs in front of the head or the trunk, the whole body region needs to be hidden, as the positioning of the hand itself (let alone its movement) might suffice to identify the sign. (Should we also find cases where signs inflectable for location in signing space need to be anonymized, that region could be shrunk down.)

Combinations of these approaches may apply.

Our experiments showed that blackening these areas is less disturbing for the viewer than a pixelation good enough to really hide the sign/mouthing.

This leaves the question where in the image the area to be covered is. In order to assist the manual annotation, our annotation environment features some computer vision algorithms, including face/mouth and hand tracking reliable enough to be used for this purpose as the areas detected need to be enlarged anyway.

The trackers generate annotation, in this case rectangle coordinates which upon export of the movie files are used to command FFmpeg (a cross-platform multimedia processing framework, cf. http://ffmpeg.org) to render the designated blocks black over the timespans specified.

In the long run, we not only want to publish front view camera perspectives, but profile views as well. Therefore, we need to make sure that the blackening can be applied to corresponding regions in the different perspectives. While face and hand tracking work sufficiently well also in the profile views, the mouth tracker needs to be reconstructed from profile face and frontal mouth tracking.

5. Conclusion

With all these parts combined, we have an equivalent to the beep found in spoken language recordings. Determining these beeps is a partially automated process that we think is good enough not to bring our informants or ourselves into trouble when publishing the corpus data.

6. Acknowledgements

This publication has been produced in the context of the joint research funding of the German Federal Government and Federal States in the Academies' Programme, with funding from the Federal Ministry of Education and Research and the Free and Hanseatic City of Hamburg. The Academies' Programme is coordinated by the Union of the German Academies of Sciences and Humanities.

7. Bibliographical References

Hinrichs, M., Zastrow, T., Hinrichs, E. (2010). WebLicht: Web-based LRT services in a distributed escience infrastructure. In *Proceedings of the seventh international conference on language resources and evaluation*. European Language Resources Association, pp. 489-493.

Hanke, T., Storz, J. (2008). iLex – A database tool for integrating sign language corpus linguistics and sign language lexicography. In *Proceedings of the 3rd workshop on the representation and processing of sign languages: construction and exploitation of sign language corpora*. European Language Resources Association, pp. 64-67.

Nishio, R., Hong, S.-E., König, S., Konrad, R., Langer, G., Hanke, T., Rathmann, C. (2010). Elicitation methods in the DGS (German Sign Language) corpus project. In *Proceedings of the 4th workshop on representation and processing of sign languages: corpora and sign language technologies*. European Language Resources Association, pp. 178-185.

Rock, F. (2001). Policy and practice in the anonymisation of linguistic data. *International Journal of Corpus Linguistics*, 6(1), pp. 1--26.

Sharoff, S. (2006). Methods and tools for development of the Russian reference corpus. In A. Wilson, D. Archer, D. & P. Rayson (Eds.), *Corpus linguistics around the world*. Amsterdam: Rodopi, pp. 167-180.