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

Julian Bleicken, Thomas Hanke, Uta Salden, Sven Wagner
University of Hamburg, Institute of German Sign Language and Communication of the Deaf

The German Sign Language Corpus

For the German Sign Language (DGS) Corpus 165 pairs of informants were filmed while conversing in DGS. The dialog formats used ranged from staged communicative events to free conversations. This resulted in 825 hours of DGS films. 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.

Parts of the material will be published with German translation only. With the annotations of these films still ongoing or not yet started, we cannot use detailed annotations in order to identify critical passages (see the box on the right).

Diverse approaches for identification of named entities are presented and evaluated on this poster:

- Using the concepts list of our database on the translations
- Having a person watch the video and tag name references
- Using named entity recognition on the translations
- Checking name lists against the translations

What has to be anonymized?

We were lucky that informants felt comfortable and partly forgot that they were filmed as this resulted in close-to-natural conversation. But in consequence informants often revealed details about themselves or other persons not really suitable to be made public, e.g.

- sequences with very private stories of the informants
- sequences with stories of wrongdoings of informants
  - ➡ cancelled for publication by informants or by the Corpus Project
- names of third parties
- names of geolocations if they could contribute to the identification of a certain person

We decided that names of third parties do not have to be anonymized in all cases. They are to be

- ➡ visible if the person is well known in the Deaf community even if the content is negative
- ➡ visible if the person is some official and the content is positive
- ➡ to be anonymized if the content refers to private life or is negative in nature

Evaluation of approaches

For this experiment different approaches to identify named entities were applied to evaluate their reliability.

The sample consisted of 31 minutes of corpus data in total from three different conversations. As we wanted different DGS dialects to be covered in the sample, we chose informants from the North as well as 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 members working in the North of Germany.

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 examined approaches. For the evaluation of this experiment, extensive additional checking of the data revealed no deficits of the so constructed ground truth.

Only for the visual inspection approach original language data in DGS was used, the other three approaches were run on the German translations. 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). 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.

Detecting names in Sign Language Corpus Data

Inspecting the FILM manually

A deaf annotator was asked to view the video and mark each occurrence of a name. The annotator had not seen the experiment data before and was allowed to stop and review the video as often as necessary. Nineteen minutes of the sample were signed in a dialect unfamiliar to the annotator. 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. The inspection of the DGS films took five times real time of the films. It is therefore a rather costly method, but it was also the most reliable method tested.

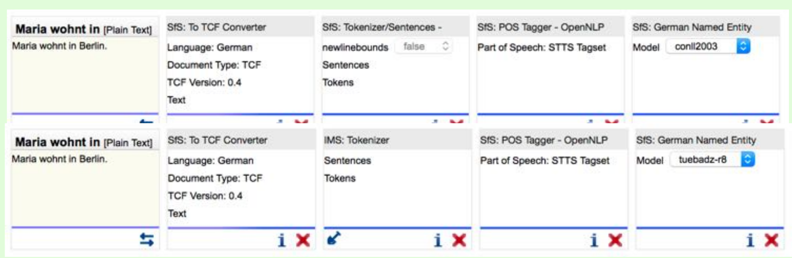
In total, there were only four false negatives, one to be neglected, because a name was mentioned in the translation only. However, one of these cases was a name that had to be anonymized. Assuming we had relied only on the manual inspection, we had missed this entity.

Named entities specifically used in the Deaf community were identified almost exclusively by the manual inspection.

true positives	93 %
false positives	5 %
false negatives	7 %

Using the TRANSLATIONS DGS-German

WebLicht



For named entity recognition, we implemented calling pre-defined WebLicht chains into our annotation environment iLex. We ran our data through different named entity recognizers available in WebLicht and finally kept working with two WebLicht chains that showed the best results in combination. 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. Surprisingly, sentences ending with an exclamation mark were not processed properly by the named entity recognizers used. After automatically adding a full stop after the exclamation marks, we got reliable results.

Half of the false negatives are due to the fact that the named entity recognizers were not run on the original language data but on the translations.

true positives	86 %
false positives	26 %
false negatives	14 %

List of names

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) were used to check our data against. Multi-part expressions were split and, where appropriate, names were extended by plural and genitive endings. Some names of the list produced too many false positives and thus were removed from the list. Nevertheless, the approach still produced a lot of false positives, requiring extra time for verification of the results.

At least in the sample at hand, the names list did not contribute any name finding that was not 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 could be added to the list.

true positives	70 %
false positives	258 %
false negatives	30 %

List of concepts

Conventionalized name signs for cities or persons well known in the Deaf community are marked as concepts in our database. From these entries a list was generated that contained the names in full length and additionally all single parts of a name, e.g. first and surname. The list was checked against the German translations, matches were considered as name reference candidates and visually inspected. It did not provide additional name references to the other approaches. However, a hit simplifies the decision whether or not a name should be anonymized.

true positives	53 %
false positives	5 %
false negatives	47 %

Anonymizing the data

For the publication of the data, named entities have to be anonymized in different places of the transcript.

In the **translations**, names are replaced by numbered placeholders, e.g. #Name1, allowing the reader to follow co-references.

In the **transcript**, the same applies to mouthing annotations and to the gloss tier in most cases.

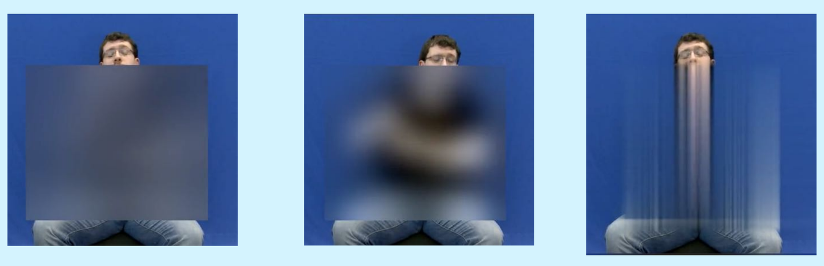
For the **video**, the time span to be anonymized is annotated separately. However, some experiments showed that completely blackening that timespan invalidates the whole sentence for further linguistic analysis as suprasegmental signals are disturbed. Therefore, we defined several options how to manipulate 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 surroundings 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.

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



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 multimedial processing framework, cf. http://ffmpeg.org) to render the designated blocks black over the timespans specified.

Conclusion

The personal inspection by a signer yielded the best results, but is rather costly. However, automatic procedures with high rates of false positives cause substantial costs for manually identifying the false alarms as such, too. Therefore, a one-pass manual inspection combined with the other automatic methods seems appropriate to gain reliable results. 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 translations.

Poster presented at the *10th Language Resources and Evaluation Conference – LREC 2016*, in Portorož, Slovenia. May 25-27, 2016.

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.