

Purpose of Corpus Anonymisation

- Ensure that no personal information is shared without informed consent
- Applies to the participants in the corpus and any third parties mentioned (consent can be obtained for participants but not for third parties)
- Informed consent is a complicated issue and varies depending on community size, corpus content and technological background

Uses of Anonymised Corpus Data

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

Finding What to Anonymise

Search for names which may identify people:

- manual inspection of videos: accurate but labour-intensive
- manual inspection of annotations e.g. glosses, translations, mouthing: accurate but labour-intensive
- Use of automatic NLP techniques to extract potential names from annotations, e.g. Named Entity Recognition on annotations: less labour-intensive but less accurate

Bleicken et al., (2016) found that a combination of automatic methods and a one-pass manual inspection was most effective

Bleicken et al. (2016). Using a Language Technology Infrastructure for German in order to Anonymize German Sign Language Corpus Data. In Proceedings of LREC 2016.

Anonymising Annotations

Replace names in annotations through:

- **Pseudonymisation** e.g. Emma -> Jenny
- **Categorisation**
 - Replace names with type tags; prevents corpus users easily searching through annotations for names to see what a signer says about someone else but loses co-reference information e.g. EMMA -> NAME, Bielefeld -> LOCATION
 - Add an index to type tags; more time-consuming but preserves co-reference information where more than one name is mentioned e.g. Alex -> NAME1, Jenny -> NAME2

Alternatives to Anonymisation

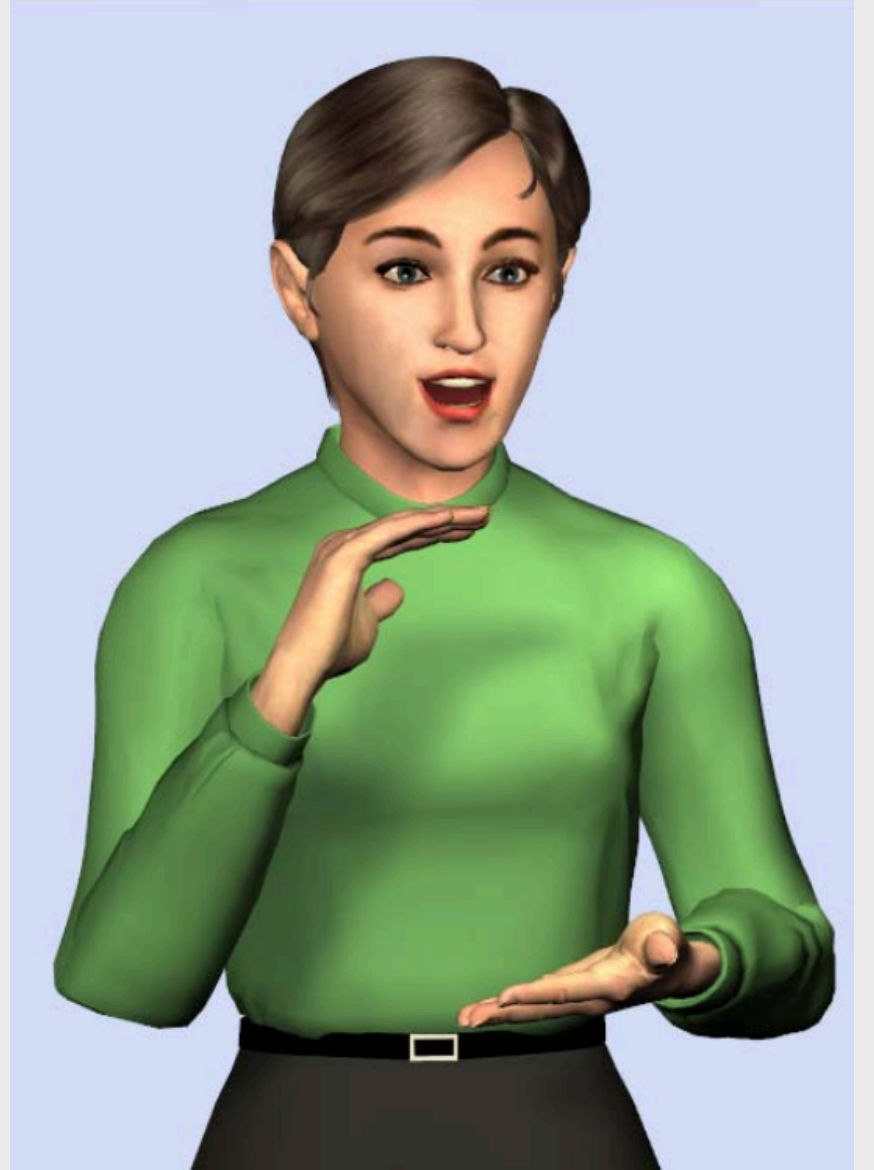
Anonymisation of an entire corpus is expensive and time-consuming. Other possibilities include:

- publicly release only parts of the data where no personal information is revealed
- ensure that informed consent has been acquired to the best standard possible (does not apply to third party information)
- ensure 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

Anonymising Video

Completely hide identity of signer

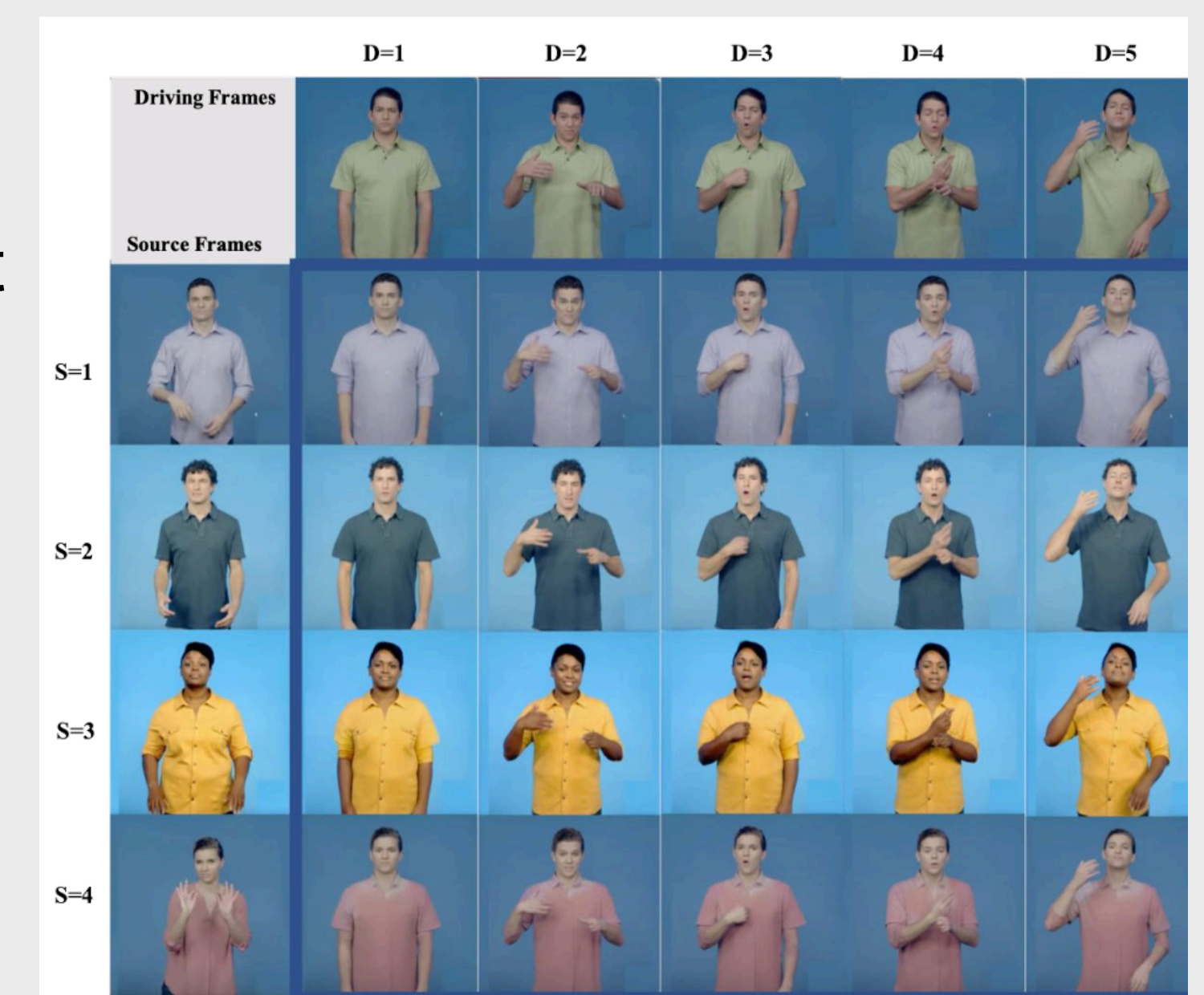
Protects visual identity of signer, but not information in utterances

- **Re-signed by actor:**
Natural look and motion, but labour-intensive; no post-processing corrections possible; potential differences in affect
- **Virtual avatar:**
Shows no real person, can be post-edited, but results depend on quality of avatar design and input
 

Paula Avatar from The American Sign Language Project
- **Motion capture input:**
Natural motion, but requires motion capture setup for original recording; signing style can hint at identity
- **Hand-animated input:**
Compositional and works without video, but often less natural; sign vocabulary labour-intensive to create; potential differences in affect

- **Video transformation:**
Natural look and motion but technology still under development and evaluation; requires high-quality training data; signing style can hint at identity

Xia et al. (2022). Sign Language Video Anonymization. In Proceedings of the 10th Workshop on the Representation and Processing of Sign Languages, LREC 2022.



Video Anonymisation Examples from Figure 6 of Xia et al., 2022.

Obscure some sensitive signs

Anonymises information in utterances, but not signer identity

Blur vs Blackening:

- Computational methods can undo some blur

How much to obscure:

- Obscuring whole image: easier, but maximum information loss
- Targeting only relevant area: more effort but less disruptive

Example below shows video anonymisation with blackening of relevant areas, and anonymisation of translation and mouthing annotations with categorisation and indices

	Translation	Lexeme/Sign	Mouth	Translation	Lexeme/Sign	Mouth
00:12:51:43	There were					
00:12:52:02	#Name2 and					
00:12:52:44	#Name3, but					
00:12:53:04	you probably	\$NAME	#name2			
00:12:53:22	don't know					
00:12:54:10	them and/	\$NAME	#name2			
00:12:54:26						
00:12:54:26						
00:12:55:35				#Name2, hm/		
00:12:55:35						
00:12:55:38						
00:12:55:41						
00:12:55:41						
00:12:56:07						
00:12:56:07						
00:12:56:27						
00:12:56:27						
00:12:57:03						
00:12:57:03						
00:12:57:33		\$LIST1:2of2d*				
00:12:57:33						
00:12:57:44						
00:12:57:44						
00:12:58:11		\$NAME	#name3			

Anonymised video and annotations from the Public DGS Corpus (Konrad et al. 2020)

Konrad et al., 2020. MY DGS – annotated. Public Corpus of German Sign Language, 3rd release [Dataset]. Universität Hamburg.