The Corpus Question

Heather Grizzle · September 27, 2026

Whose signing builds sign language AI, what they agreed to, what they were paid, and whether their faces can be withdrawn: the corpus question buyers inherit.

目录

Every sign language AI system is built from recordings of people signing. Whose signing it was, what they agreed to, what they were paid, and whether they can ever take their faces back are questions the market has not answered and most buyers have not asked.

IN SHORT

- A signed language is produced on the face and body, so a sign language training corpus is a collection of identifiable recordings of people from a small community. It cannot be anonymised without destroying the linguistic content it exists to capture.
- The corpora the field documents were assembled in four ways: scraped from public video, taken from broadcast interpretation, recorded in studios with recruited signers, or collected privately by vendors. Each carries a different answer to consent, payment, authorship and withdrawal, and only the first three are documented at all.
- The public research corpora carry use restrictions (academic and computational use only, non-commercial, research-only data-sharing agreements) that a commercial product would have to honour. Whether deployed systems honour them is not visible from the outside.
- Biometric privacy law turns on identification. A face processed for its grammar rather than its identity can fall outside the protections written for faces.
- A buyer inherits the corpus. Five questions, put in writing before contract, establish whether a vendor can account for the people its system was built from.

The face is the data

In spoken and written language technology, the training data and the person can usually be separated. A sentence can be stripped of its author. A voice can be transformed until the speaker is unrecognisable and the words remain. Signed languages do not permit this separation, because the language is produced on the body. The hands carry the lexicon, but the grammar is distributed across the face: eyebrow position marks questions, mouth patterns disambiguate signs, head tilt and eye gaze establish reference and agreement. A recording of a signer with the face blurred is not a de-identified language sample. It is a corrupted one.

This has a consequence that the machine learning literature has recognised for some time and the procurement literature has not. A sign language corpus is, necessarily, a

collection of identifiable recordings of specific people. The signing community in any country is small, and the interdisciplinary review of sign language AI datasets led by Bragg and colleagues observed that "because of the small size of the deaf community, it may be possible to personally identify signers in videos"[1]. A model trained on such a corpus has learned from these people's faces. If the system generates signing, it has learned to produce faces that resemble theirs. The question of whose signing built the system is therefore not an abstract question of data ethics. It is a question about named, findable individuals whose likeness is now part of a product.

Where the corpora come from

The documented corpora that underpin published research in sign language recognition and generation were assembled in a small number of ways, and the assembly method determines nearly everything that follows.

> A recording of a signer with the face blurred is not a de-identified language sample. It is a corrupted one.

The first method is collection from public video. One widely used word-level American Sign Language dataset was compiled by downloading videos from YouTube and sign language teaching sites, organised into roughly two thousand glosses, and released under Microsoft's Computational Use of Data Agreement with the statement that all the data "is intended for academic and computational use only" and that "no commercial usage is allowed"[2]. Its documentation records a signer identifier for each clip but contains no statement about the signers' consent, fluency or awareness that their video had been collected. Subsequent linguistic examination found its English gloss labels "pervasive[ly]" inconsistent, with the same sign labelled by different words and the same word applied to different signs[3]. A larger corpus released in 2023 gathered 984 hours of footage from more than 2,500 signers by retrieving public YouTube videos tagged as sign language content; its authors explain that they included "videos across all skill levels and signing styles, as long as they were comprehensible to an ASL user", that they release only video identifiers rather than the videos themselves, and that a creator's route to withdrawal is to make the video private or delete it[4].

The second method is broadcast interpretation. The largest British Sign Language corpus consists of roughly 1,467 hours of BBC programmes carrying an in-vision interpreter, produced by 39 interpreters in total, and is available for non-commercial research under a data-sharing agreement with the broadcaster that obliges researchers to delete portions or the whole of the dataset if instructed[5]. The German weather-broadcast corpus that anchored the first neural sign language translation results was built the same way, from television interpreters rendering a hearing presenter's script[6]. These interpreters were paid to interpret a broadcast. They were not paid to supply a training corpus, and the language they produced is interpretation under time pressure from a spoken source. De Sisto and colleagues warn that using "non-native and/or interpreted data to train translation systems that should be able to recognise, and generate, spontaneous authentic/native language will lead to less natural and very likely inaccurate translations", and describe the result as a form of translationese in which the signed stream "will always be somehow affected by the source spoken language"[7].

The third method is studio recording with recruited signers. The best-known example is a corpus of more than eighty hours of continuous American Sign Language recorded in controlled conditions with consenting participants and released under a Creative Commons licence that forbids commercial use[8]. This is the model that looks most like ordinary research ethics: participants recruited, informed, compensated as research participants, and recorded for a stated purpose. It is also the most expensive way to build a corpus, which is why it produces the smallest ones.

Figure 1. Four documented research corpora, by size and by the terms attached to them. Vendor training data, the corpus that matters most to a buyer, appears in no dataset paper and carries no licence page.
Novara Consulting Group

The fourth method is private collection by the companies that sell sign language AI products. This is the method that matters most for a buyer, and it is the one about

which the least is documented. A vendor's own training data does not appear in a dataset paper. It has no licence page. Whether it was scraped, purchased, recorded with hired signers, contributed by users of the product, or assembled from the public research corpora described above, with or without regard to their use restrictions, is not visible from outside the company.

Four questions the assembly method answers

Set beside each other, the documented corpora answer four questions in four different ways, and the differences are the whole subject.

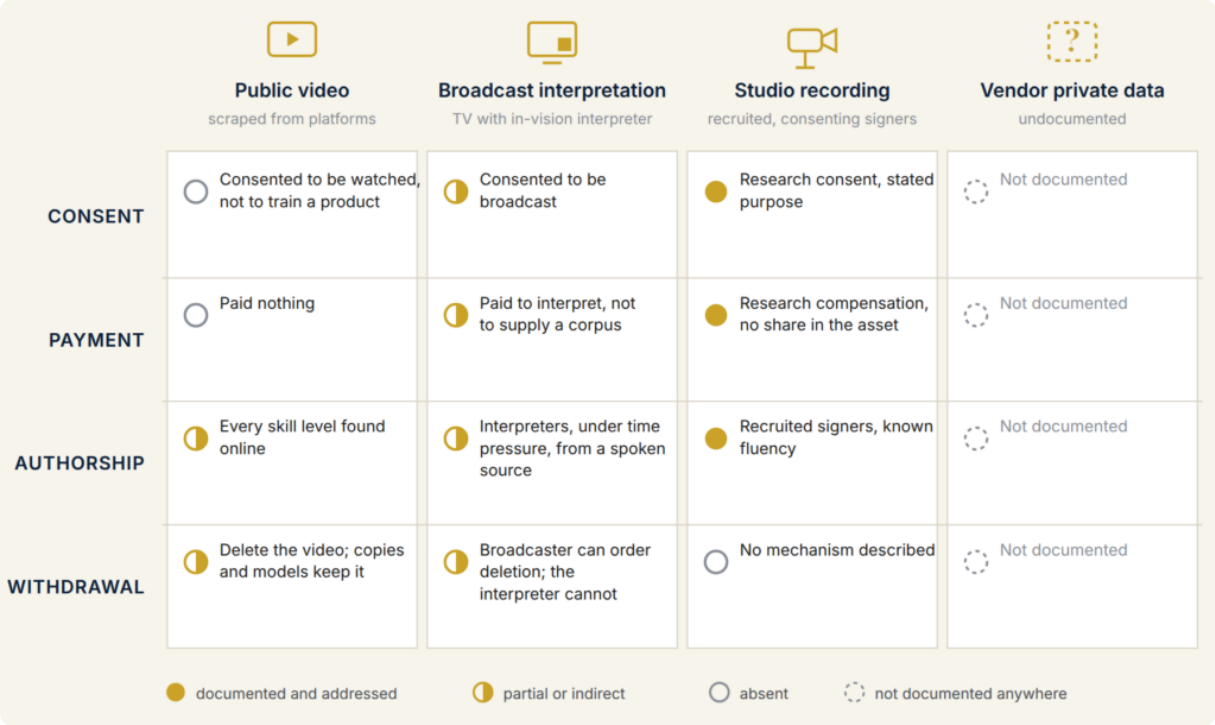

Figure 2. What each assembly method answers on consent, payment, authorship and withdrawal. The fourth column is the one a deployed product is most likely to have been built from.
Novara Consulting Group

The first question is consent. A person who uploaded a video to teach a sign consented to being watched. Whether that consent extends to being decomposed into training features for a commercial product is a question the platform's terms of service answer in the platform's favour and the person's own understanding almost certainly does not. Broadcast interpreters consented to be broadcast. Studio participants consented to research. In none of the documented cases outside the studio corpora is there

evidence that the signer was asked whether their signing could be used to build a machine that signs.

In none of the documented cases outside the studio corpora is there evidence that the signer was asked whether their signing could be used to build a machine that signs.

The second question is payment. The people whose signing produced the largest corpora were paid nothing for it, because the corpora were built from work they had already done for other reasons or for no reason at all. The interpreters were paid by broadcasters for interpreting. The uploaders were paid by no one. Only the studio participants received anything, and they received research compensation, not a share in a commercial asset. Bragg and colleagues note that data agreements "may specify compensation for usage, describe the allowed uses of the data" and set time limits on use[1]. The documented public corpora do not do this because they were not built on agreements with signers. They were built on agreements with platforms and broadcasters.

The third question is authorship, in the sense of whose language the system has learned. A corpus of interpreters has learned interpretation, with its lag, its source-language shadow and its professional register. A corpus of every skill level found on a video platform has learned a mixture of native signing, second-language signing and learner signing, weighted by whatever people happened to upload. Bragg and colleagues note that "the majority of sign language users may not be deaf, and first-language signers only make up a small fraction of users for most sign languages"[1], and a corpus assembled from public video inherits that proportion. A Deaf-led review of 101 sign language AI papers found the field "dominated by hearing non-signing researchers" and "driven by what decisions are the most convenient or perceived as important to hearing researchers"[9]. Convenience is also what decides which corpus gets used. The question of whose language a system speaks is settled, in practice, by which recordings were easiest to obtain.

The fourth question is withdrawal. A person who appears in a dataset assembled from public video can delete the video. If the dataset was distributed as video identifiers,

the deletion propagates to anyone who has not already downloaded the clip. It does not propagate to anyone who has, and it does not propagate into a model that was trained before the deletion. The broadcast corpus is the only documented case with a contractual deletion obligation, and that obligation runs from the broadcaster to the researcher, not from the interpreter to anyone. A trained model has no mechanism for forgetting one signer. Withdrawal, in the sense a person would understand it, is not available anywhere in the documented landscape once training has happened.

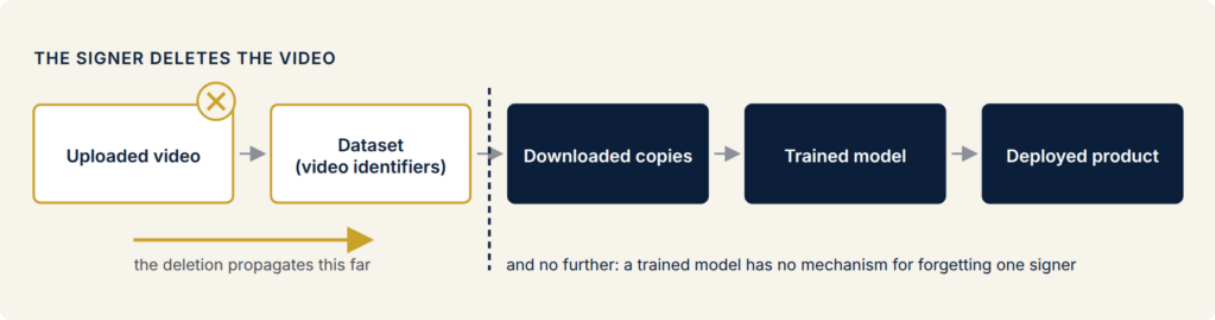

Figure 3. Withdrawal in the documented landscape. Deleting a public video reaches a dataset distributed as identifiers and stops there.
Novara Consulting Group

The gap in the law

Readers who assume that faces are protected by biometric privacy law should look at how that law is drafted. The Illinois Biometric Information Privacy Act, the strictest of the American statutes, defines a biometric identifier as "a retina or iris scan, fingerprint, voiceprint, or scan of hand or face geometry" and expressly excludes photographs from the definition; it requires a written release before a private entity collects such an identifier[10]. Under the European and United Kingdom data protection regimes, biometric data becomes special category data when it is processed "for the purpose of uniquely identifying a natural person", and the United Kingdom's regulator is explicit that ordinary photographs and video of people are "not automatically biometric data even if you use it for identification purposes"; the special category status attaches when specific technical processing produces a template for matching[11].

A sign language model processes the face in extraordinary detail. It extracts eyebrow position, mouth shape, gaze direction and head movement frame by frame, because those features are the grammar. But it does not process them to identify the signer. It

processes them to learn the language. On the face of the statutes, that is exactly the purpose the biometric categories were not written for. The most intensive computational processing of a person's face that exists in ordinary commerce may therefore sit outside the protections that were drafted with faces in mind, because the drafters assumed a face would only be processed to say who it belongs to. This is not a settled legal conclusion and Novara Consulting Group offers no legal opinion on any jurisdiction. It is an observation that the categories do not fit, and that a vendor asked whether its corpus is "biometric data" may be able to answer no in perfectly good faith while holding thousands of hours of identifiable Deaf faces.

The licences are a separate and simpler matter. The word-level corpus permits academic and computational use only. The studio corpus is non-commercial. The broadcast corpus is research-only under an agreement that forbids cooperation with commercial partners[7]. A commercial product trained on any of them would be operating outside its data's terms. No published mechanism exists for a buyer to find out whether that has happened, and a buyer reading vendor documentation will rarely find a statement of corpus provenance detailed enough to rule it out.

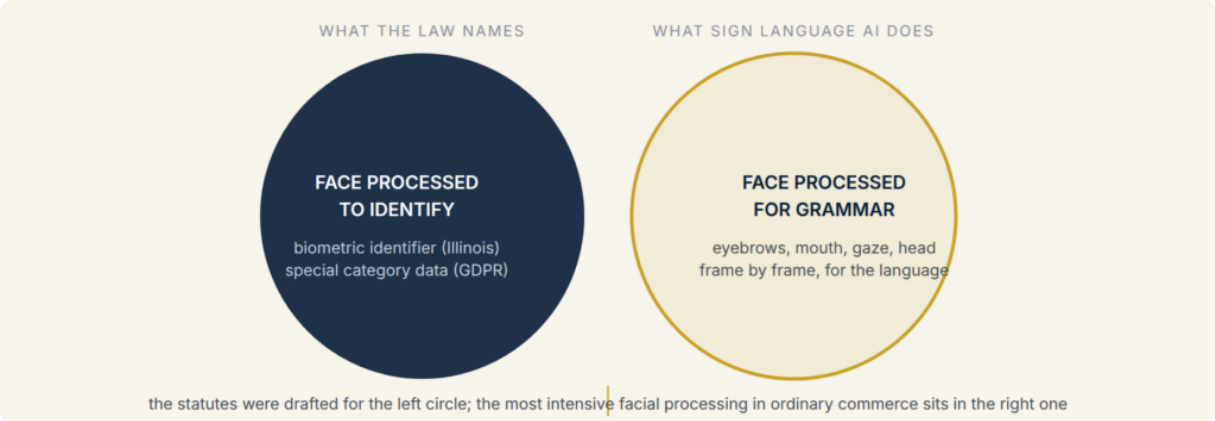

Figure 4. Biometric privacy categories attach to processing for identification. Sign language AI processes the face for its grammar.
Novara Consulting Group

What a buyer inherits

An institution that deploys a sign language AI system inherits its corpus. It inherits the consent that was or was not obtained, the licence terms that were or were not honoured, the interpreters' register or the learners' errors, and the impossibility of withdrawal. It inherits them in a specific form: the Deaf people it serves will be

watching a system produce faces learned from members of their own community, and some of those people will be able to recognise whose face it is.

This is why the corpus question precedes the accuracy question. A system whose output is excellent but whose corpus was taken without consent has not solved a governance problem; it has monetised one. A system whose corpus is fully accounted for but whose output is poor is at least a system whose failures can be discussed with the people it affects. The Convention on the Rights of Persons with Disabilities requires States Parties, in developing policies that concern persons with disabilities, to "closely consult with and actively involve" them through their representative organisations[12]. A corpus assembled from people who were never asked is the plainest available example of a technology developed without that involvement, and an institution that buys it is not consulting anyone either.

> **FIVE QUESTIONS TO PUT TO A VENDOR IN WRITING**
>
> First, what are the sources of the training data, by category and proportion: public video, broadcast, studio recording, user contributions, licensed third-party corpora? Second, for each source, what did the signers agree to, in what document, and does that document mention commercial use and generation? Third, what were the signers paid, by whom, and for what? Fourth, which public research corpora were used, and under which licence terms is commercial deployment permitted? Fifth, what happens when a signer asks to be removed, and does the answer include the trained model or only the stored video? A vendor that can answer these in writing has a data governance function. A vendor that cannot has a corpus.

The answers will not, in most cases, be reassuring. That is not a reason to leave the questions unasked. It is the reason to ask them before signature, when the answers still change the decision, rather than after deployment, when they change only the liability. The buyers who ask first will set the terms on which the rest of the market is eventually required to answer.

参考文献

1. Bragg, D., Caselli, N., Hochgesang, J. A., Huenerfauth, M., Katz-Hernandez, L., Koller, O., Kushalnagar, R., Vogler, C., and Ladner, R. E. (2021). The FATE landscape of sign language AI datasets: An interdisciplinary perspective. ACM Transactions on Accessible Computing, 14(2), Article 7. https://doi.org/10.1145/3436996
2. Li, D., Rodriguez, C., Yu, X., and Li, H. (2020). Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Dataset documentation and licence statement at https://github.com/dxli94/WLASL
3. Neidle, C., and Ballard, C. (2022). Why alternative gloss labels will increase the value of the WLASL dataset. American Sign Language Linguistic Research Project Report No. 21, Boston University. https://www.bu.edu/asllrp/rpt21/asllrp21.pdf
4. Uthus, D., Tanzer, G., and Georg, M. (2023). YouTube-ASL: A large-scale, open-domain American Sign Language-English parallel corpus. Advances in Neural Information Processing Systems 36, Datasets and Benchmarks Track.
5. Albanie, S., Varol, G., Momeni, L., Bull, H., Afouras, T., Chowdhury, H., Fox, N., Woll, B., Cooper, R., McParland, A., and Zisserman, A. (2021). BBC-Oxford British Sign Language dataset. arXiv:2111.03635. Dataset terms at https://www.robots.ox.ac.uk/~vgg/data/bobsl/
6. Camgöz, N. C., Hadfield, S., Koller, O., Ney, H., and Bowden, R. (2018). Neural sign language translation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 7784-7793.
7. De Sisto, M., Vandeghinste, V., Egea Gómez, S., De Coster, M., Shterionov, D., and Saggion, H. (2022). Challenges with sign language datasets for sign language recognition and translation. Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC 2022), 2478-2487. https://aclanthology.org/2022.lrec-1.264

8. Duarte, A., Palaskar, S., Ventura, L., Ghadiyaram, D., DeHaan, K., Metze, F., Torres, J., and Giró-i-Nieto, X. (2021). How2Sign: A large-scale multimodal dataset for continuous American Sign Language. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Licence statement at https://how2sign.github.io/
9. Desai, A., De Meulder, M., Hochgesang, J. A., Kocab, A., and Lu, A. X. (2024). Systemic biases in sign language AI research: A Deaf-led call to reevaluate research agendas. arXiv:2403.02563.
10. Illinois Biometric Information Privacy Act, 740 ILCS 14/10 (definitions) and 14/15(b) (consent). https://law.justia.com/codes/illinois/chapter-740/act-740-ilcs-14/
11. Information Commissioner's Office (United Kingdom). What is special category data? Guidance on Article 9 UK GDPR, biometric data. https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/lawful-basis/special-category-data/what-is-special-category-data/
12. United Nations (2006). Convention on the Rights of Persons with Disabilities, Article 4(3). https://www.un.org/development/desa/disabilities/convention-on-the-rights-of-persons-with-disabilities.html

---

引用方式

Grizzle, H. M. (2026, September 27). The Corpus Question. Novara Consulting Group. https://doi.org/10.5281/zenodo.23003735