NOVARA
CONSULTING GROUP

This document was translated from English by machine and may contain mistakes. The English original is at https://www.novaracg.com/2026/09/27/the-corpus-question/.

La question du corpus

Heather Grizzle · September 27, 2026

Qui produit la langue des signes utilisée pour entraîner l'IA, ce à quoi ces personnes ont consenti, ce qu'elles ont été payées, et si leurs visages peuvent être retirés : la question du corpus dont héritent les acheteurs.

Sommaire

1. Le visage, c'est la donnée
2. D'où viennent les corpus
3. Quatre questions auxquelles répond la méthode de constitution
4. La faille du droit
5. Ce dont hérite un acheteur
6. Références

Chaque système d'IA en langue des signes est construit à partir d'enregistrements de personnes en train de signer. De qui était ce signe, ce que ces personnes ont accepté, ce qu'elles ont été payées, et si elles peuvent un jour reprendre leur visage, sont des questions auxquelles le marché n'a pas répondu et que la plupart des acheteurs n'ont pas posées.

EN BREF

- Une langue des signes se produit sur le visage et le corps, si bien qu'un corpus d'entraînement en langue des signes est une collection d'enregistrements identifiables de personnes issues d'une petite communauté. Il ne peut être anonymisé sans détruire le contenu linguistique qu'il a pour vocation de capturer.
- Les corpus que le domaine documente ont été constitués de quatre manières : extraits de vidéos publiques, tirés d'interprétations diffusées, enregistrés en studio avec des signeurs recrutés, ou collectés en privé par des fournisseurs. Chacun apporte une réponse différente aux questions de consentement, de rémunération, de paternité et de retrait, et seules les trois premières sont documentées de quelque manière que ce soit.
- Les corpus de recherche publics comportent des restrictions d'usage (usage académique et computationnel uniquement, non commercial, accords de partage de données réservés à la recherche) qu'un produit commercial devrait respecter. Que les systèmes déployés les respectent ou non n'est pas visible de l'extérieur.
- Le droit de la vie privée biométrique repose sur l'identification. Un visage traité pour sa grammaire plutôt que pour son identité peut échapper aux protections conçues pour les visages.
- Un acheteur hérite du corpus. Cinq questions, posées par écrit avant tout contrat, permettent d'établir si un fournisseur peut rendre compte des personnes à partir desquelles son système a été construit.

Le visage, c'est la donnée

Dans les technologies du langage parlé et écrit, les données d'entraînement et la personne peuvent généralement être dissociées. Une phrase peut être dépouillée de son auteur. Une voix peut être transformée jusqu'à devenir méconnaissable tandis que les mots demeurent. Les langues des signes ne permettent pas cette dissociation, car la langue se produit sur le corps. Les mains portent le lexique, mais la grammaire se distribue sur le visage : la position des sourcils marque les questions, les mouvements de la bouche désambiguïsent les signes, l'inclinaison de la tête et la direction du regard

établissent la référence et l'accord. Un enregistrement d'un signeur au visage flouté n'est pas un échantillon linguistique désidentifié. C'en est un corrompu.

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.

D'où viennent les corpus

Les corpus documentés qui sous-tendent la recherche publiée en reconnaissance et génération de langue des signes ont été constitués selon un petit nombre de méthodes, et la méthode de constitution détermine presque tout ce qui suit.

Un enregistrement d'un signeur au visage flouté n'est pas un échantillon linguistique désidentifié. C'en est un corrompu.

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].

La troisième méthode est l'enregistrement en studio avec des signeurs recrutés. L'exemple le plus connu est un corpus de plus de quatre-vingts heures d'ASL continu enregistré dans des conditions contrôlées avec des participants consentants, diffusé sous une licence Creative Commons qui interdit tout usage commercial[8]. C'est le modèle qui ressemble le plus à une éthique de recherche ordinaire : des participants recrutés, informés, rémunérés en tant que participants de recherche, et enregistrés dans un but déclaré. C'est aussi la manière la plus coûteuse de constituer un corpus, ce qui explique qu'elle en produise les plus petits.



Figure 1. Quatre corpus de recherche documentés, par taille et par conditions qui leur sont attachées. Les données d'entraînement des fournisseurs, le corpus qui compte le plus pour un acheteur, n'apparaissent dans aucun article de présentation d'ensemble de données et ne portent aucune page de licence.
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.

Quatre questions auxquelles répond la méthode de constitution

Mis côte à côte, les corpus documentés répondent à quatre questions de quatre manières différentes, et ces différences constituent l'ensemble du sujet.

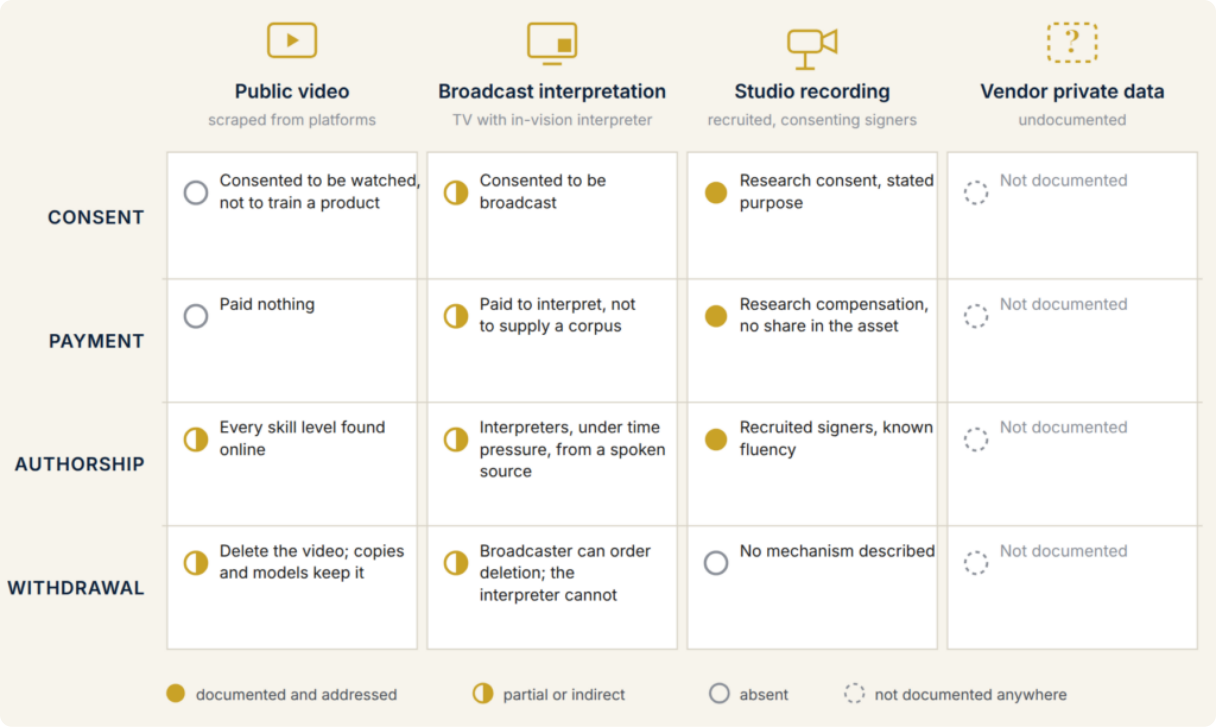


Figure 2. Ce que chaque méthode de constitution répond en matière de consentement, de rémunération, de paternité et de retrait. La quatrième colonne est celle à partir de laquelle un produit déployé a le plus de chances d'avoir été construit.
Novara Consulting Group

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.

Dans aucun des cas documentés en dehors des corpus de studio n'existe de preuve que le signeur ait été inter-

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

rogé sur la possibilité que son signe serve à construire une machine qui signe.

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.

La quatrième question est celle du retrait. Une personne qui apparaît dans un ensemble de données constitué à partir de vidéos publiques peut supprimer la vidéo. Si l'ensemble de données a été diffusé sous forme d'identifiants vidéo, la suppression se propage à quiconque n'a pas déjà téléchargé le clip. Elle ne se propage pas à ceux qui l'ont fait, et elle ne se propage pas dans un modèle qui a été entraîné avant la suppression. Le corpus de diffusion est le seul cas documenté comportant une obligation contractuelle de suppression, et cette obligation va du diffuseur au chercheur, non de l'interprète à qui que ce soit. Un modèle entraîné ne dispose d'aucun mécanisme permettant d'oublier un signeur en particulier. Le retrait, au sens où une personne l'entendrait, n'est disponible nulle part dans le paysage documenté une fois l'entraînement effectué.

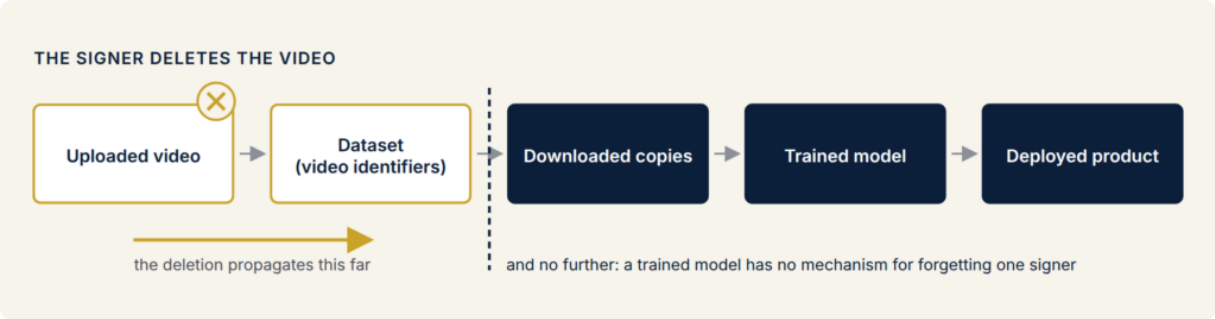


Figure 3. Le retrait dans le paysage documenté. Supprimer une vidéo publique atteint un ensemble de données diffusé sous forme d'identifiants, et s'arrête là.
Novara Consulting Group

La faille du droit

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.

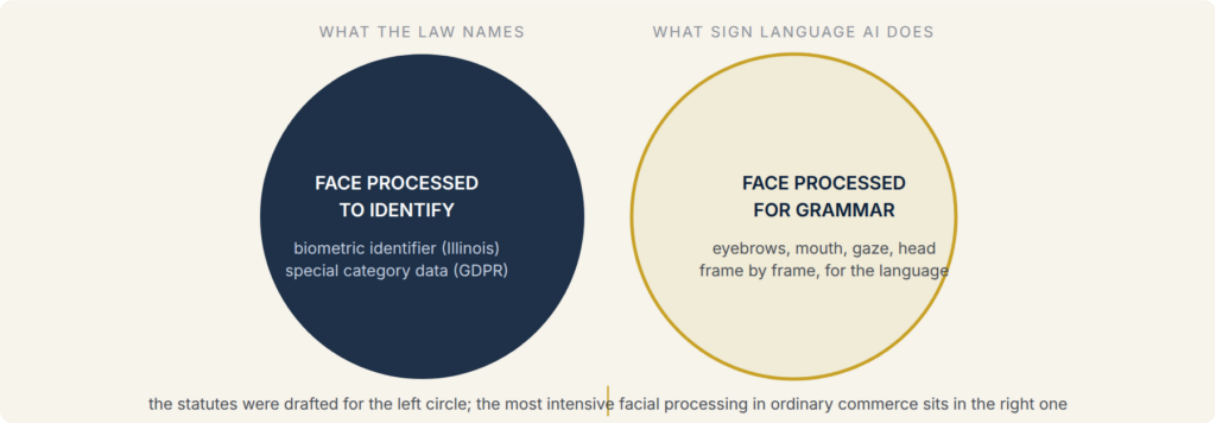


Figure 4. Les catégories de protection biométrique s'attachent au traitement à des fins d'identification. L'IA en langue des signes traite le visage pour sa grammaire.
Novara Consulting Group

Ce dont hérite un acheteur

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 organisa-

tions[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.

> **CINQ QUESTIONS À POSER PAR ÉCRIT À UN FOURNISSEUR**
>
> Premièrement, quelles sont les sources des données d'entraînement, par catégorie et par proportion : vidéo publique, diffusion, enregistrement en studio, contributions d'utilisateurs, corpus tiers sous licence ? Deuxièmement, pour chaque source, à quoi les signeurs ont-ils consenti, dans quel document, et ce document mentionne-t-il l'usage commercial et la génération ? Troisièmement, qu'ont été payés les signeurs, par qui, et pour quoi ? Quatrièmement, quels corpus de recherche publics ont été utilisés, et selon quelles conditions de licence le déploiement commercial est-il autorisé ? Cinquièmement, que se passe-t-il lorsqu'un signeur demande à être retiré, et la réponse inclut-elle le modèle entraîné ou seulement la vidéo stockée ? Un fournisseur capable de répondre par écrit à ces questions dispose d'une fonction de gouvernance des données. Un fournisseur qui ne le peut pas dispose d'un corpus.

Les réponses ne seront, dans la plupart des cas, pas rassurantes. Ce n'est pas une raison pour ne pas poser les questions. C'est la raison de les poser avant la signature, quand les réponses peuvent encore changer la décision, plutôt qu'après le déploiement, quand elles ne changent plus que la responsabilité. Les acheteurs qui posent ces questions en premier fixeront les conditions auxquelles le reste du marché devra tôt ou tard répondre.

Références

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

TIPS

Know something about a sign language AI system that the public should know? Send it to tips@novaracg.com. Every tip is read by me and kept confidential. If you work in the UK, writing to us is not by itself a protected disclosure under the Public Interest Disclosure Act 1998, so before you share employer information, check your options with Protect (protect-advice.org.uk), the free whistleblowing advice line.

Comment citer

Grizzle, H. M. (2026, September 27). La question du corpus. Novara Consulting Group. https://doi.org/10.5281/zenodo.23003735