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 cuestión del corpus

Heather Grizzle · 27 de septiembre de 2026

De quién es el lenguaje de señas que alimenta la IA de lenguaje de señas, qué acordaron, cuánto se les pagó y si sus rostros pueden retirarse: la cuestión del corpus que heredan los compradores.

Contenidos

1. El rostro es el dato
2. De dónde provienen los corpus
3. Cuatro preguntas que responde el método de ensamblaje
4. El vacío en la ley
5. Lo que hereda un comprador
6. Referencias

Todo sistema de IA de lenguaje de señas se construye a partir de grabaciones de personas que hacen señas. De quién era ese lenguaje de señas, qué acordaron, cuánto se les pagó y si alguna vez pueden recuperar sus rostros son preguntas que el mercado no ha respondido y que la mayoría de los compradores no ha formulado.

EN RESUMEN

- Una lengua de señas se produce con el rostro y el cuerpo, de modo que un corpus de entrenamiento de lengua de señas es una colección de grabaciones identificables de personas de una comunidad pequeña. No puede anonimizarse sin destruir el contenido lingüístico que existe para capturar.
- Los corpus que documenta el campo se ensamblaron de cuatro maneras: extraídos de video público, tomados de interpretación en transmisiones, grabados en estudios con firmantes reclutados, o recopilados de forma privada por proveedores. Cada uno conlleva una respuesta distinta sobre consentimiento, pago, autoría y retiro, y solo los primeros tres están documentados de alguna forma.
- Los corpus públicos de investigación llevan restricciones de uso (uso académico y computacional únicamente, no comercial, acuerdos de intercambio de datos solo para investigación) que un producto comercial tendría que respetar. Si los sistemas desplegados los respetan no es visible desde fuera.
- La ley de privacidad biométrica gira en torno a la identificación. Un rostro procesado por su gramática en lugar de su identidad puede quedar fuera de las protecciones redactadas para los rostros.
- Un comprador hereda el corpus. Cinco preguntas, formuladas por escrito antes del contrato, establecen si un proveedor puede dar cuenta de las personas a partir de las cuales se construyó su sistema.

El rostro es el dato

En la tecnología de lenguaje hablado y escrito, los datos de entrenamiento y la persona por lo general pueden separarse. Una frase puede despojarse de su autor. Una voz puede transformarse hasta que el hablante resulte irreconocible y las palabras permanezcan. Las lenguas de señas no permiten esta separación, porque la lengua se produce sobre el cuerpo. Las manos portan el léxico, pero la gramática se distribuye por el rostro: la posición de las cejas marca las preguntas, los patrones de la boca desambiguan señas, la inclinación de la cabeza y la dirección de la mirada establecen referencia y concordancia. Una grabación de un firmante con el rostro difuminado no es una muestra de lenguaje anonimizada. Es una muestra corrupta.

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.

De dónde provienen los corpus

Los corpus documentados que sustentan la investigación publicada en reconocimiento y generación de lenguaje de señas se ensamblaron mediante un pequeño número de métodos, y el método de ensamblaje determina casi todo lo que sigue.

> Una grabación de un firmante con el rostro difuminado no es una muestra de lenguaje anonimizada. Es una muestra corrupta.

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].

El tercer método es la grabación en estudio con firmantes reclutados. El ejemplo más conocido es un corpus de más de ochenta horas de lengua de señas americana continua, grabado en condiciones controladas con participantes que dieron su consentimiento, publicado bajo una licencia Creative Commons que prohíbe el uso comercial[8]. Este es el modelo que más se parece a la ética de investigación ordinaria: participantes reclutados, informados, compensados como participantes de investigación, y grabados con un propósito declarado. También es la forma más costosa de construir un corpus, razón por la cual produce los más pequeños.



Figura 1. Cuatro corpus documentados de investigación, por tamaño y por los términos que llevan asociados. Los datos de entrenamiento del proveedor, el corpus que más importa a un comprador, no aparece en ningún artículo sobre conjuntos de datos ni lleva página de licencia.
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.

Cuatro preguntas que responde el método de ensamblaje

Puestos uno junto a otro, los corpus documentados responden cuatro preguntas de cuatro maneras distintas, y las diferencias son el tema central.

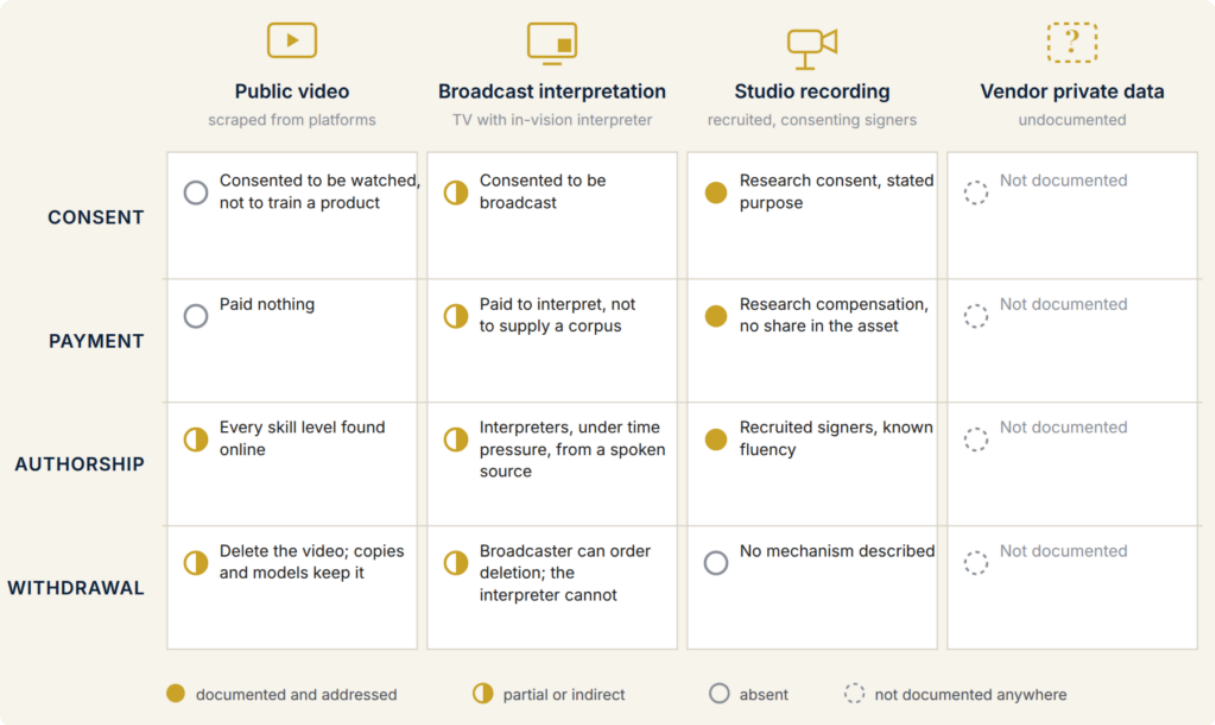


Figura 2. Lo que responde cada método de ensamblaje sobre consentimiento, pago, autoría y retiro. La cuarta columna es la más probable a partir de la cual se haya construido un producto desplegado. Novara Consulting Group

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.

En ninguno de los casos documentados fuera de los corpus de estudio hay evidencia de que se le preguntara al firmante si su lenguaje de señas podía usarse para construir una máquina que hace señas.

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.

La cuarta pregunta es el retiro. Una persona que aparece en un conjunto de datos ensamblado a partir de video público puede eliminar el video. Si el conjunto de datos se distribuyó como identificadores de video, la eliminación se propaga a cualquiera que aún no haya descargado el clip. No se propaga a quien ya lo hizo, ni se propaga a un modelo que fue entrenado antes de la eliminación. El corpus de transmisión es el único caso documentado con una obligación contractual de eliminación, y esa obligación va de la emisora al investigador, no del intérprete a nadie. Un modelo entrenado no tiene ningún mecanismo para olvidar a un firmante. El retiro, en el sentido que una persona lo entendería, no está disponible en ninguna parte del panorama documentado una vez que ha ocurrido el entrenamiento.

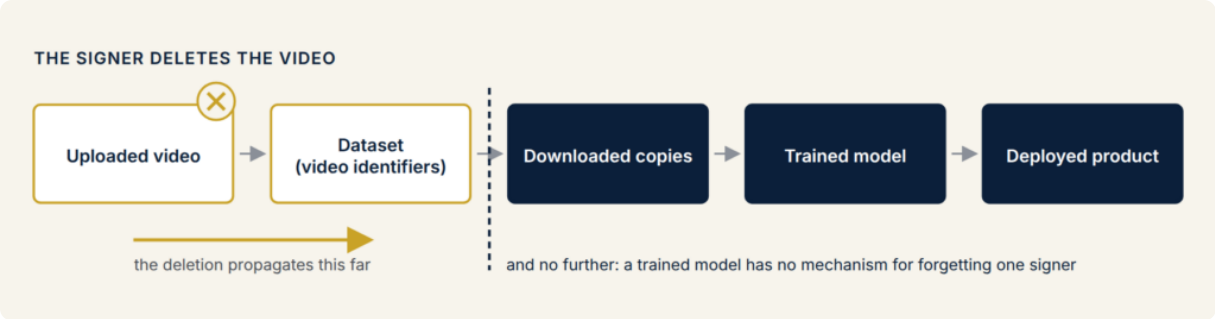


Figura 3. El retiro en el panorama documentado. Eliminar un video público llega a un conjunto de datos distribuido como identificadores, y ahí se detiene.
Novara Consulting Group

El vacío en la ley

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.

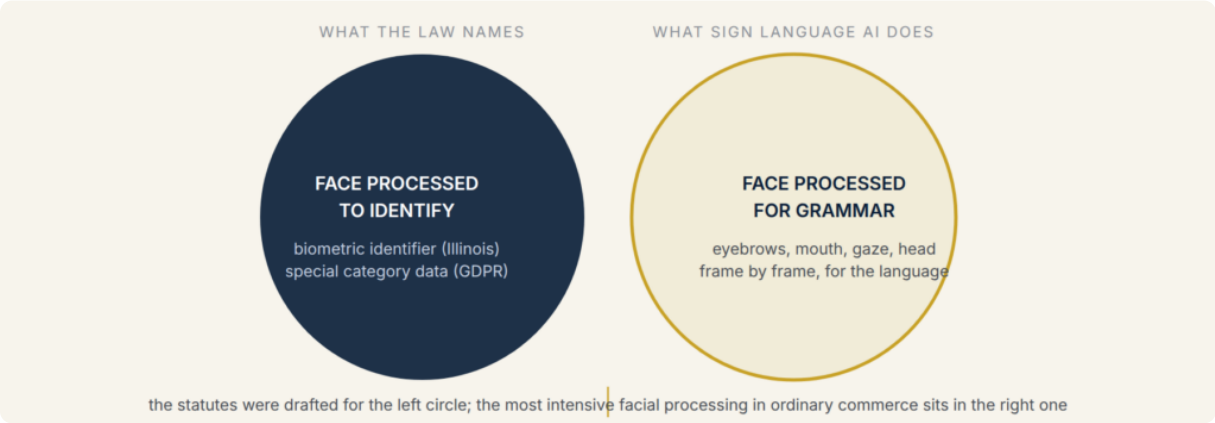


Figura 4. Las categorías de privacidad biométrica se aplican al procesamiento para identificación. La IA de lenguaje de señas procesa el rostro por su gramática.
Novara Consulting Group

Lo que hereda un comprador

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.

> **CINCO PREGUNTAS QUE PLANTEAR A UN PROVEEDOR POR ESCRITO**
>
> Primero, ¿cuáles son las fuentes de los datos de entrenamiento, por categoría y proporción: video público, transmisión, grabación en estudio, aportes de usuarios, corpus de terceros con licencia? Segundo, para cada fuente, ¿qué aceptaron los firmantes, en qué documento, y ese documento menciona el uso comercial y la generación? Tercero, ¿qué se les pagó a los firmantes, por parte de quién, y por qué? Cuarto, ¿qué corpus públicos de investigación se usaron, y bajo qué términos de licencia se permite el despliegue comercial? Quinto, ¿qué sucede cuando un firmante pide ser eliminado, y la respuesta incluye el modelo entrenado o solo el video almacenado? Un proveedor que pueda responder esto por escrito tiene una función de gobernanza de datos. Un proveedor que no puede tiene un corpus.

Las respuestas no serán, en la mayoría de los casos, tranquilizadoras. Eso no es razón para dejar las preguntas sin formular. Es la razón para formularlas antes de la firma,

cuando las respuestas aún cambian la decisión, en lugar de después del despliegue, cuando solo cambian la responsabilidad. Los compradores que preguntan primero establecerán los términos bajo los cuales el resto del mercado eventualmente se verá obligado a responder.

Referencias

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 langua-

ge 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.

Cómo citar

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