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 questione del corpus

Heather Grizzle · September 27, 2026

Chi firma la lingua dei segni che alimenta l'IA, a cosa ha acconsentito, quanto è stato pagato e se i volti possono essere ritirati: la questione del corpus che gli acquirenti ereditano.

Contenuti

1. Il volto è il dato
2. Da dove provengono i corpora
3. Quattro domande a cui risponde il metodo di assemblaggio
4. Il vuoto nella legge
5. Cosa eredita un acquirente
6. Riferimenti

Ogni sistema di intelligenza artificiale per la lingua dei segni è costruito a partire da registrazioni di persone che segnano. Chi fossero queste persone, cosa abbiano accettato, quanto siano state pagate e se possano mai riottenere i propri volti sono domande a cui il mercato non ha risposto e che la maggior parte degli acquirenti non si è mai posta.

IN BREVE

- Una lingua dei segni si produce sul volto e sul corpo, quindi un corpus di addestramento per la lingua dei segni è una raccolta di registrazioni identificabili di persone appartenenti a una piccola comunità. Non può essere reso anonimo senza distruggere il contenuto linguistico che esiste per catturare.
- I corpora documentati nel settore sono stati assemblati in quattro modi: estratti da video pubblici, ricavati da interpretazioni televisive, registrati in studio con segnanti reclutati, oppure raccolti privatamente da fornitori. Ciascuno comporta una risposta diversa in materia di consenso, pagamento, paternità e possibilità di ritiro, e solo i primi tre sono in qualche modo documentati.
- I corpora di ricerca pubblici comportano restrizioni d'uso (uso accademico e computazionale esclusivo, non commerciale, accordi di condivisione dati solo per la ricerca) che un prodotto commerciale dovrebbe rispettare. Se i sistemi effettivamente distribuiti li rispettino non è verificabile dall'esterno.
- Il diritto sulla privacy biometrica si basa sull'identificazione. Un volto elaborato per la sua grammatica anziché per la sua identità può restare al di fuori delle tutele previste per i volti.
- Un acquirente eredita il corpus. Cinque domande, messe per iscritto prima del contratto, permettono di stabilire se un fornitore è in grado di rendere conto delle persone a partire dalle quali è stato costruito il suo sistema.

Il volto è il dato

Nelle tecnologie per il linguaggio parlato e scritto, i dati di addestramento e la persona possono di solito essere separati. Una frase può essere privata del suo autore. Una voce può essere trasformata fino a renderla irriconoscibile, mentre le parole restano. Le lingue dei segni non consentono questa separazione, perché la lingua si produce sul corpo. Le mani portano il lessico, ma la grammatica è distribuita sul volto: la posizione delle sopracciglia segna le domande, i pattern della bocca disambiguano i segni, l'inclinazione della testa e la direzione dello sguardo stabiliscono riferimento e accordo. Una registrazione di un segnante con il volto sfocato non è un campione linguistico de-identificato. È un campione corrotto.

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.

Da dove provengono i corpora

I corpora documentati alla base della ricerca pubblicata sul riconoscimento e la generazione della lingua dei segni sono stati assemblati con un numero ristretto di metodi, e il metodo di assemblaggio determina quasi tutto ciò che segue.

> Una registrazione di un segnante con il volto sfocato non è un campione linguistico de-identificato. È un campione corrotto.

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].

Il terzo metodo è la registrazione in studio con segnanti reclutati. L'esempio più noto è un corpus di oltre ottanta ore di lingua dei segni americana continua, registrato in condizioni controllate con partecipanti consenzienti e rilasciato con licenza Creative Commons che vieta l'uso commerciale[8]. Questo è il modello che più si avvicina all'etica della ricerca ordinaria: partecipanti reclutati, informati, compensati in quanto partecipanti alla ricerca, e registrati per uno scopo dichiarato. È anche il modo più costoso di costruire un corpus, motivo per cui produce i corpora più piccoli.



Figura 1. Quattro corpora di ricerca documentati, per dimensione e per i termini ad essi associati. I dati di addestramento dei fornitori, il corpus che più conta per un acquirente, non compare in nessun articolo sui dataset e non ha alcuna pagina di licenza.
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.

Quattro domande a cui risponde il metodo di assemblaggio

Messi a confronto, i corpora documentati rispondono a quattro domande in quattro modi diversi, e le differenze sono l'intero oggetto della questione.

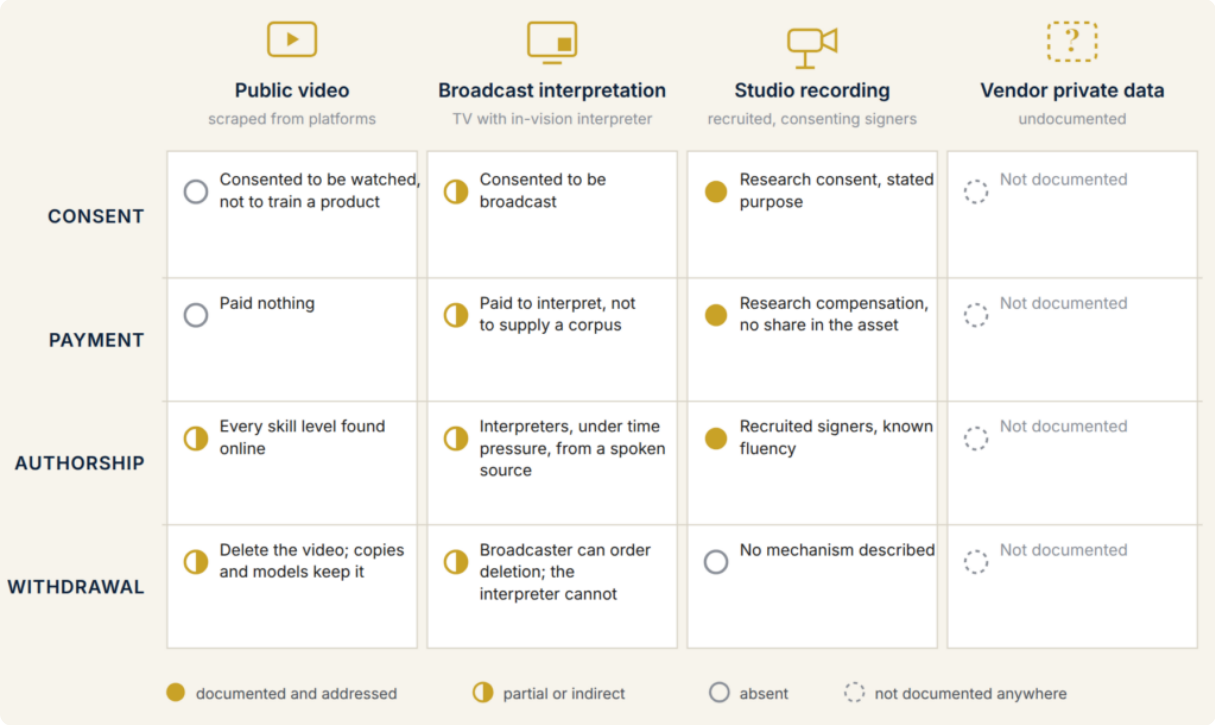


Figura 2. Cosa risponde ciascun metodo di assemblaggio in materia di consenso, pagamento, paternità e ritiro. La quarta colonna è quella da cui un prodotto distribuito ha più probabilmente avuto origine.
Novara Consulting Group

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 nessuno dei casi documentati al di fuori dei corpora di studio vi è prova che al segnante sia stato chiesto se il suo segnare potesse essere usato per costruire una macchina che segna.

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 quarta domanda riguarda il ritiro. Una persona che compare in un dataset assemblato da video pubblici può eliminare il video. Se il dataset è stato distribuito come identificativi di video, l'eliminazione si propaga a chiunque non abbia già scaricato la clip. Non si propaga a chi l'ha già scaricata, e non si propaga in un modello che è stato addestrato prima dell'eliminazione. Il corpus televisivo è l'unico caso documentato con un obbligo contrattuale di eliminazione, e tale obbligo va dall'emittente al ricercatore, non dall'interprete a chiunque altro. Un modello addestrato non ha alcun meccanismo per dimenticare un singolo segnante. Il ritiro, nel senso in cui una persona lo intenderebbe, non è disponibile in nessun punto del panorama documentato una volta che l'addestramento ha avuto luogo.

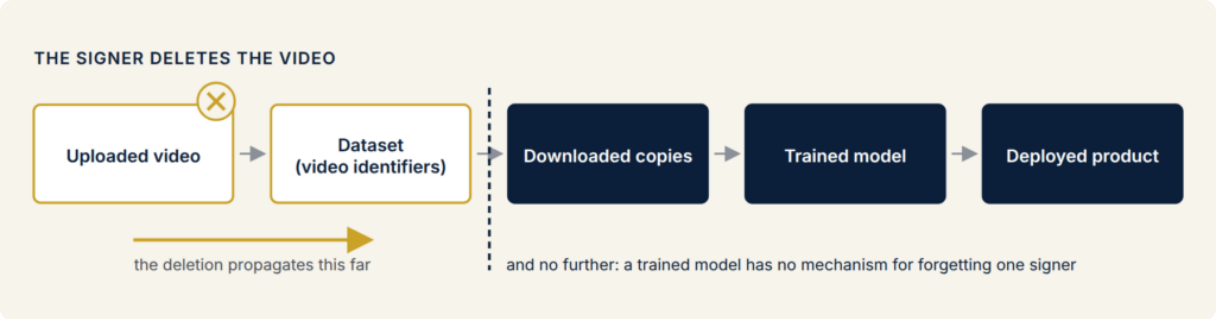


Figura 3. Il ritiro nel panorama documentato. L'eliminazione di un video pubblico raggiunge un dataset distribuito come identificativi, e lì si ferma.
Novara Consulting Group

Il vuoto nella legge

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.

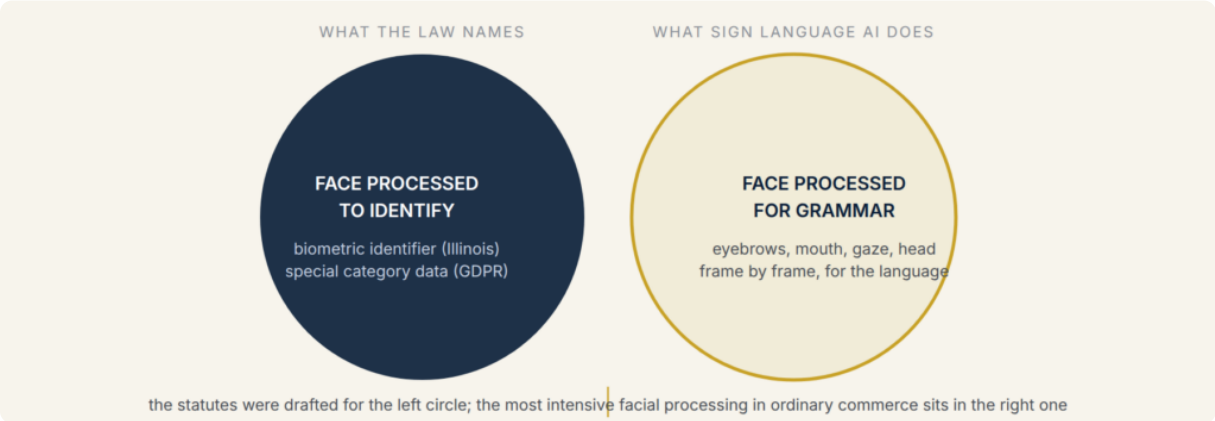


Figura 4. Le categorie di privacy biometrica si applicano al trattamento a fini identificativi. L'intelligenza artificiale per la lingua dei segni elabora il volto per la sua grammatica.
Novara Consulting Group

Cosa eredita un acquirente

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.

> **CINQUE DOMANDE DA PORRE PER ISCRITTO A UN FORNITORE**
>
> Primo, quali sono le fonti dei dati di addestramento, per categoria e proporzione: video pubblici, trasmissioni, registrazioni in studio, contributi degli utenti, corpora di terze parti su licenza? Secondo, per ciascuna fonte, a cosa hanno acconsentito i segnanti, in quale documento, e tale documento menziona l′uso commerciale e la generazione? Terzo, quanto sono stati pagati i segnanti, da chi, e per cosa? Quarto, quali corpora di ricerca pubblici sono stati utilizzati, e secondo quali termini di licenza è consentita la distribuzione commerciale? Quinto, cosa accade quando un segnante chiede di essere rimosso, e la risposta include il modello addestrato o solo il video archiviato? Un fornitore in grado di rispondere per iscritto a queste domande dispone di una funzione di governance dei dati. Un fornitore che non può farlo dispone solo di un corpus.

Le risposte non saranno, nella maggior parte dei casi, rassicuranti. Questo non è un motivo per lasciare le domande inespresse. È il motivo per porle prima della firma,

quando le risposte possono ancora cambiare la decisione, piuttosto che dopo la distribuzione, quando cambiano solo la responsabilità. Gli acquirenti che le pongono per primi stabiliranno i termini a cui il resto del mercato sarà infine tenuto a rispondere.

Riferimenti

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 Resour-

ces 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.

Come citare

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