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/.

A Questão do Corpus

Heather Grizzle · September 27, 2026

Quem assina constrói a IA de língua gestual, a que é que concordaram, quanto é que lhes foi pago, e se os seus rostos podem ser retirados: a questão do corpus que os compradores herdam.

Conteúdo

1. O rosto é o dado
2. De onde vêm os corpora
3. Quatro perguntas que o método de reunião responde
4. A lacuna na lei
5. O que um comprador herda
6. Referências

Todo sistema de IA de língua gestual é construído a partir de gravações de pessoas a gesticular. De quem era a gestualização, o que essas pessoas aceitaram, quanto lhes foi pago e se alguma vez poderão retirar os seus rostos são questões que o mercado não respondeu e que a maioria dos compradores não colocou.

EM SUMA

- Uma língua gestual é produzida no rosto e no corpo, pelo que um corpus de treino de língua gestual é uma coleção de gravações identificáveis de pessoas de uma pequena comunidade. Não pode ser anonimizado sem destruir o conteúdo linguístico que existe para capturar.
- Os corpora que o campo documenta foram reunidos de quatro formas: extraídos de vídeo público, retirados de interpretação televisiva, gravados em estúdio com gesticuladores recrutados, ou recolhidos privadamente por fornecedores. Cada um carrega uma resposta diferente às questões de consentimento, pagamento, autoria e retirada, e apenas as três primeiras estão sequer documentadas.
- Os corpora de investigação pública trazem restrições de utilização (uso académico e computacional apenas, não comercial, acordos de partilha de dados apenas para investigação) que um produto comercial teria de respeitar. Se os sistemas implementados os respeitam ou não, não é visível a partir de fora.
- A lei da privacidade biométrica assenta na identificação. Um rosto processado pela sua gramática, e não pela sua identidade, pode ficar fora das proteções escritas para rostos.
- Um comprador herda o corpus. Cinco perguntas, feitas por escrito antes do contrato, estabelecem se um fornecedor consegue dar conta das pessoas a partir de quem o seu sistema foi construído.

O rosto é o dado

Na tecnologia de linguagem falada e escrita, os dados de treino e a pessoa podem geralmente ser separados. Uma frase pode ser despojada do seu autor. Uma voz pode ser transformada até o locutor se tornar irreconhecível e as palavras permanecerem. As línguas gestuais não permitem esta separação, porque a língua é produzida no corpo. As mãos transportam o léxico, mas a gramática está distribuída pelo rosto: a posição das sobrancelhas marca perguntas, os padrões da boca desambiguam sinais, a inclinação da cabeça e a direção do olhar estabelecem referência e concordância. Uma gravação de um gesticulador com o rosto desfocado não é uma amostra linguística desidentificada. É uma amostra corrompida.

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 onde vêm os corpora

Os corpora documentados que sustentam a investigação publicada em reconhecimento e geração de língua gestual foram reunidos por um pequeno número de métodos, e o método de reunião determina quase tudo o que se segue.

> Uma gravação de um gesticulador com o rosto desfocado não é uma amostra linguística desidentificada. É uma amostra corrompida.

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].

O terceiro método é a gravação em estúdio com gesticuladores recrutados. O exemplo mais conhecido é um corpus com mais de oitenta horas de Língua Gestual Americana contínua, gravado em condições controladas com participantes consentidos e disponibilizado ao abrigo de uma licença Creative Commons que proíbe o uso comercial[8]. Este é o modelo que mais se assemelha à ética de investigação comum: participantes recrutados, informados, remunerados como participantes de investigação e gravados para um fim declarado. É também a forma mais cara de construir um corpus, razão pela qual produz os mais pequenos.



Figura 1. Quatro corpora de investigação documentados, por tamanho e pelos termos a eles associados. Os dados de treino de fornecedores, o corpus que mais importa a um comprador, não aparece em nenhum artigo de conjunto de dados e não tem qualquer página de licença.
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.

Quatro perguntas que o método de reunião responde

Postos lado a lado, os corpora documentados respondem a quatro perguntas de quatro formas diferentes, e as diferenças são todo o assunto.

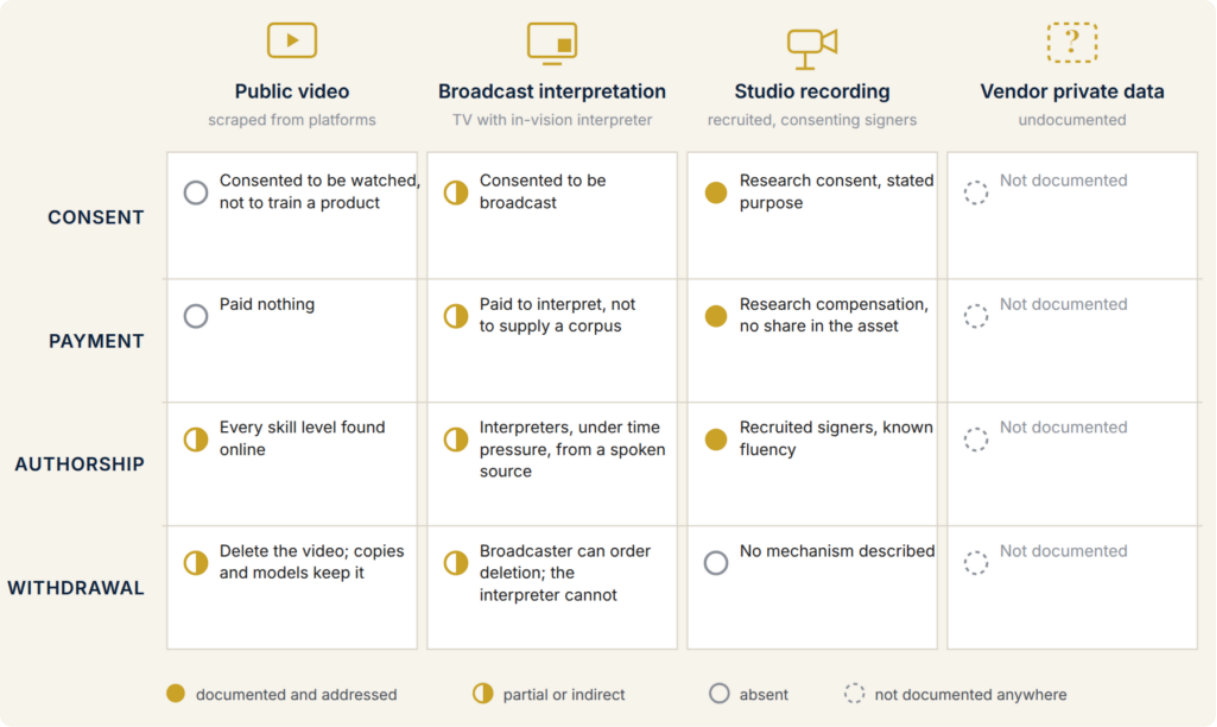


Figura 2. O que cada método de reunião responde sobre consentimento, pagamento, autoria e retirada. A quarta coluna é aquela a partir da qual é mais provável que um produto implementado tenha sido construído.
Novara Consulting Group

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.

Em nenhum dos casos documentados fora dos corpora de estúdio há provas de que ao gesticulador foi perguntado

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

se a sua gesticulação poderia ser usada para construir uma máquina que gesticula.

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.

A quarta pergunta é a retirada. Uma pessoa que aparece num conjunto de dados reunido a partir de vídeo público pode eliminar o vídeo. Se o conjunto de dados foi distribuído como identificadores de vídeo, a eliminação propaga-se para quem ainda não tenha descarregado o clipe. Não se propaga para quem já o tenha feito, nem se propaga para um modelo que foi treinado antes da eliminação. O corpus televisivo é o único caso documentado com uma obrigação contratual de eliminação, e essa obrigação vai da estação de televisão para o investigador, não do intérprete para ninguém. Um modelo treinado não tem qualquer mecanismo para esquecer um gesticulador. A retirada, no sentido em que uma pessoa a entenderia, não está disponível em lado nenhum do panorama documentado uma vez que o treino tenha ocorrido.

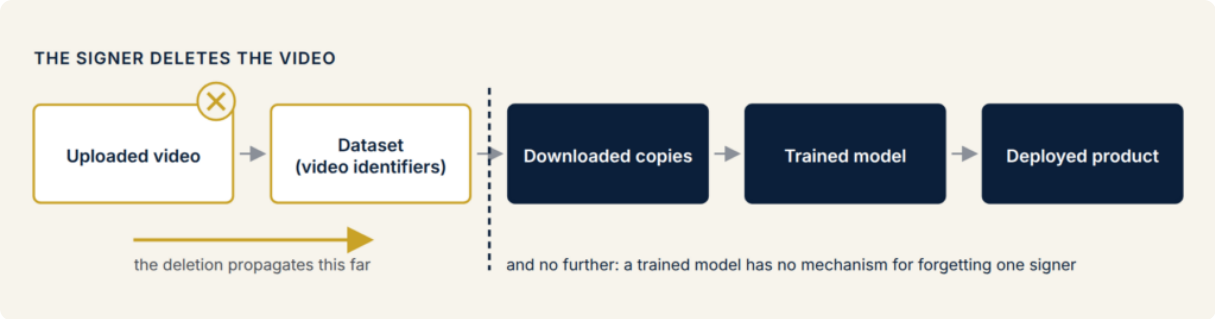


Figura 3. A retirada no panorama documentado. Eliminar um vídeo público alcança um conjunto de dados distribuído como identificadores e para aí mesmo.
Novara Consulting Group

A lacuna na lei

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.

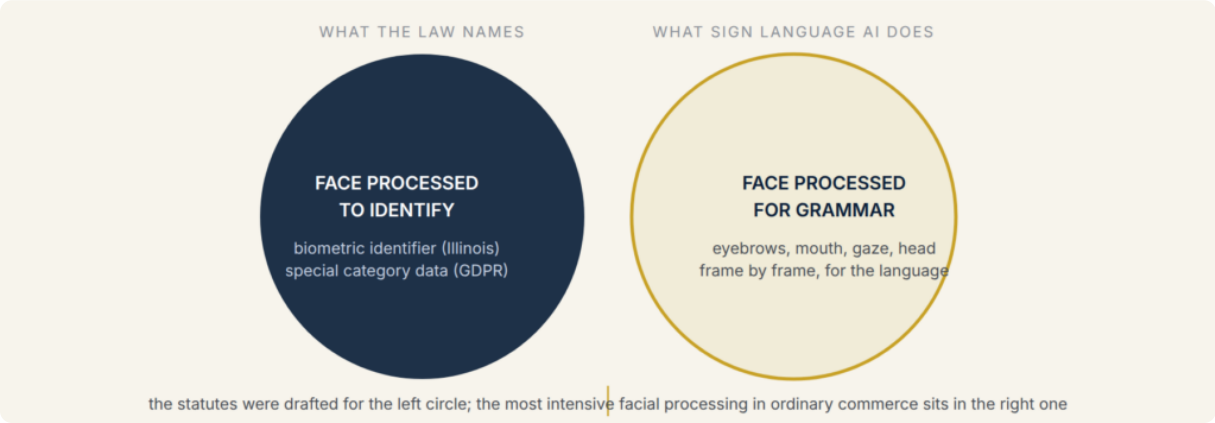


Figura 4. As categorias de privacidade biométrica aplicam-se ao processamento para identificação. A IA de língua gestual processa o rosto pela sua gramática.
Novara Consulting Group

O que um comprador herda

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 organisati-

ons[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 PERGUNTAS A COLOCAR A UM FORNECEDOR POR ESCRITO**
>
> Primeiro, quais são as fontes dos dados de treino, por categoria e proporção: vídeo público, transmissão televisiva, gravação em estúdio, contribuições de utilizadores, corpora licenciados de terceiros? Segundo, para cada fonte, a que é que os gesticuladores acederam, em que documento, e esse documento menciona o uso comercial e a geração? Terceiro, quanto foi pago aos gesticuladores, por quem, e para quê? Quarto, quais corpora de investigação pública foram usados, e sob que termos de licença é permitida a implementação comercial? Quinto, o que acontece quando um gesticulador pede para ser removido, e a resposta inclui o modelo treinado ou apenas o vídeo armazenado? Um fornecedor que consiga responder a isto por escrito tem uma função de governação de dados. Um fornecedor que não consiga tem um corpus.

As respostas não serão, na maioria dos casos, tranquilizadoras. Isso não é razão para deixar as perguntas por fazer. É a razão para as fazer antes da assinatura, quando as respostas ainda mudam a decisão, em vez de depois da implementação, quando apenas mudam a responsabilidade. Os compradores que perguntarem primeiro estabelecerão os termos em que o resto do mercado será eventualmente obrigado a responder.

Referências

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.

Como citar

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