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/.

语料问题

Heather Grizzle · September 27, 2026

谁的手语用于构建手语人工智能、他们同意了什么、他们获得了多少报酬,以及他们的面部影像能否被撤回——这是买家所继承的语料库问题。

目录

1. 面孔即数据
2. 语料库从何而来
3. 汇编方法回答的四个问题
4. 法律中的空白
5. 买家继承了什么
6. 参考文献

每一个手语人工智能系统都建立在人们打手语的录像基础之上。这些录像里的人是谁、他们同意了什么、他们得到了什么报酬、他们是否能够收回自己的面孔——这些问题,市场尚未作答,而大多数买家甚至从未提出过。

简要说明

- 手语是通过面部和身体来表达的,因此手语训练语料库是一个来自小型社群、可识别身份人员的录像集合。它无法在不破坏其所要捕捉的语言内容的前提下实现匿名化。
- 该领域记录在案的语料库是以四种方式汇编而成的:从公开视频中抓取、取自广播口译、在录音棚中由招募的打手语者录制,或由供应商私下收集。每一种方式对同意、报酬、署名权和撤回权都给出了不同的答案,而且只有前三种方式有据可查。
- 公开的研究语料库都附带使用限制(仅限学术和计算用途、非商业性、仅限研究的数据共享协议),商业产品理应遵守这些限制。至于已部署的系统是否真的遵守了这些限制,外界无从得知。
- 生物识别隐私法的核心在于身份识别。一张为了捕捉语法而非身份而被处理的面孔,可能会落在为面孔所写的保护范围之外。
- 买家继承的是语料库本身。五个问题,在签约前以书面形式提出,可以厘清供应商是否能够交代清楚其系统所依赖的那些人。

面孔即数据

在口语和书面语言技术中,训练数据与个人通常是可以分离的。一句话可以剥离其作者身份。一段声音可以被转换到说话者无法被辨认的程度,而话语本身依然完整。手语不允许这种分离,因为这种语言是通过身体来产生的。手部承载词汇,但语法则分布在面部:眉毛的位置标记疑问句,嘴型消除手势的歧义,头部倾斜和眼神注视确立指称和一致关系。一段面部被打了马赛克的打手语者录像,并不是一个去标识化的语言样本,而是一个被破坏了的样本。

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.

语料库从何而来

支撑已发表的手语识别与生成研究的、有据可查的语料库,是通过少数几种方式汇编而成的,而汇编方法几乎决定了后续的一切。

一段面部被打了马赛克的打手语者录像,并不是一个去标识化的语言样本,而是一个被破坏了的样本。

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].

第三种方法是在录音棚中由招募的打手语者录制。最著名的例子是一个超过八十小时的连续美国手语语料库,它是在受控条件下由同意参与的参与者录制而成,并依据一份禁止商业用途的知识共享许可协议发布[8]。这种模式最接近普通的研究伦理:招募参与者、告知情况、作为研究参与者给予补偿,并为既定目的进行录制。这也是构建语料库中成本最高的方式,这正是为什么由此产生的语料库规模最小。



图1。四个有据可查的研究语料库,按规模及附带条款列出。而对买家而言最重要的语料库——供应商的训练数据——却未出现在任何数据集论文中,也没有任何许可页面。
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.

汇编方法回答的四个问题

将这些有据可查的语料库并列比较,可以看到它们以四种不同的方式回答了四个问题,而这些差异正是全部问题所在。

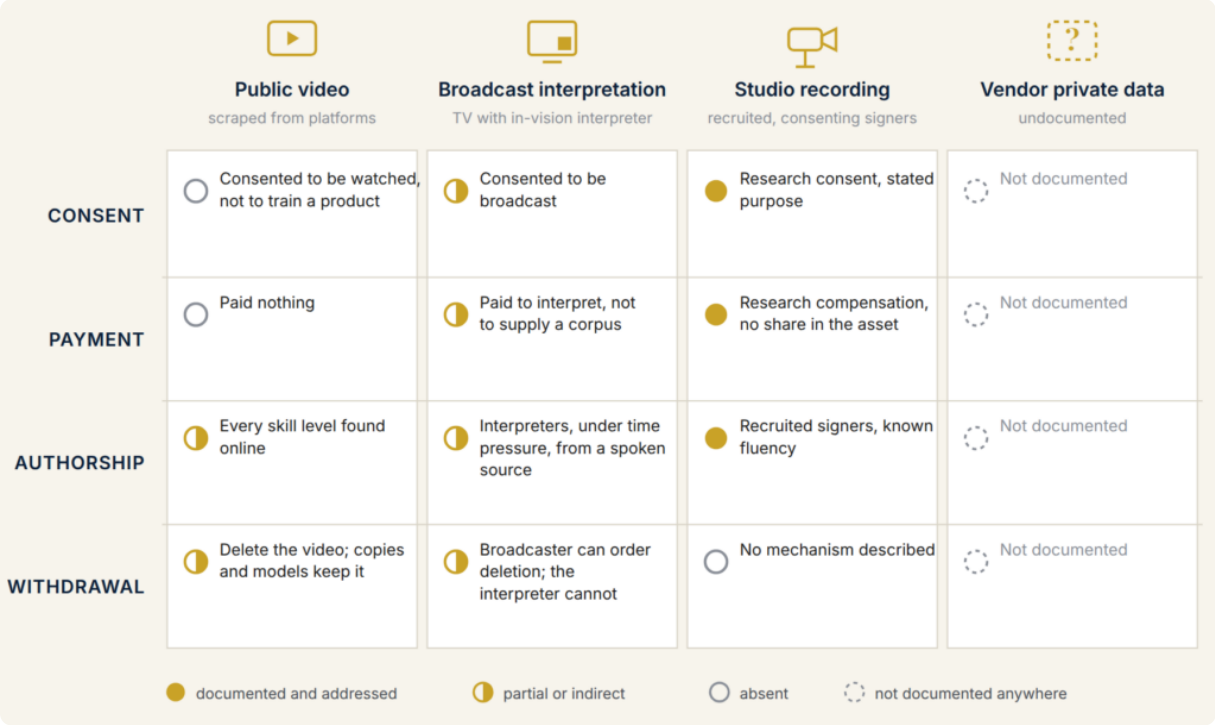


图2。每种汇编方法对同意、报酬、署名权和撤回权分别给出了怎样的答案。第四列最有可能正是一个已部署产品的构建来源。
Novara Consulting Group

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.

在录音棚语料库之外的

The second question is payment. The people whose signing produced the largest corpora were paid nothing for it, because the corpora

所有有据可查的案例中,没有证据表明打手语者曾被问及其手语是否可以被用来构建一台会打手语的机器。

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.

第四个问题是撤回。出现在由公开视频汇编而成的数据集中的人,可以删除该视频。如果该数据集是以视频标识符的形式发布的,那么删除操作会传播给尚未下载该片段的任何人,但不会传播给已经下载过的人,也不会传播进入在删除之前就已训练完成的模型。广播语料库是唯一有据可查的、附带合同删除义务的案例,而且该义务是从广播机构流向研究者,而非从译员流向任何人。一个已训练完成的模型没有任何机制可以忘记某一位打手语者。就一个人所理解的意义而言,在训练已经发生之后,撤回在整个有据可查的格局中根本不可获得。

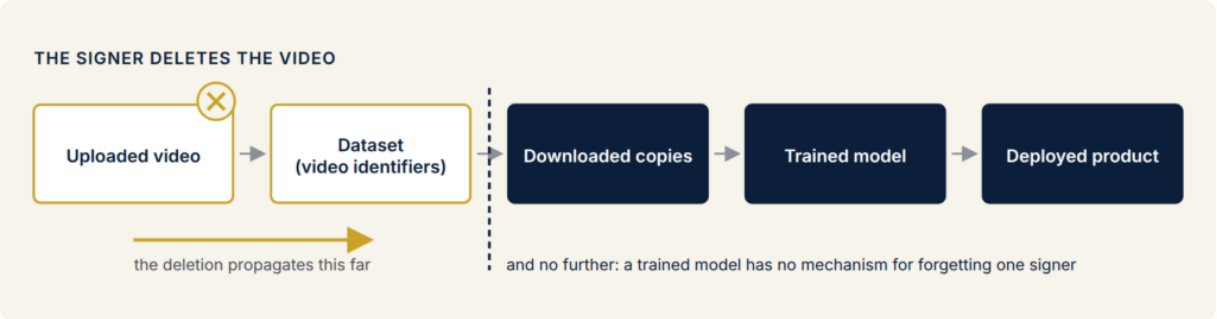


图3。有据可查格局中的撤回情况。删除一段公开视频,其影响可以触及以标识符形式发布的数据集,但也仅止于此。
Novara Consulting Group

法律中的空白

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.

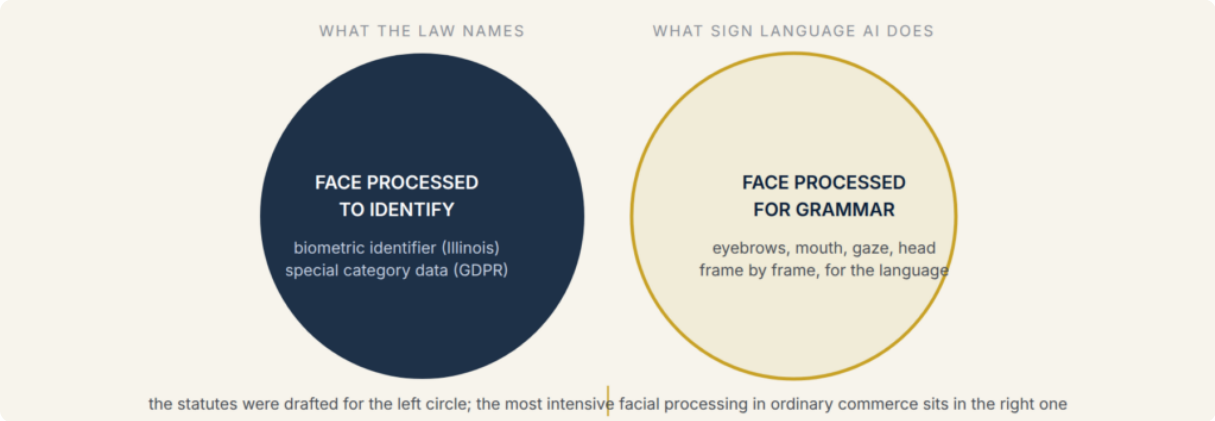


图4。生物识别隐私类别所对应的是为身份识别目的而进行的处理。而手语人工智能处理面部,是为了捕捉其语法。
Novara Consulting Group

买家继承了什么

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.

> **五个应以书面形式向供应商提出的问题**
>
> 第一,训练数据的来源是什么,按类别和比例划分:公开视频、广播、录音棚录制、用户贡献、许可的第三方语料库?第二,对于每一种来源,打手语者同意了什么内容,以何种文件形式记录,该文件是否提及商业用途和生成用途?第三,打手语者获得了什么报酬,由谁支付,因何而支付?第四,使用了哪些公开研究语料库,依据哪些许可条款允许商业部署?第五,当打手语者要求移除时会发生什么,答案是否包括已训练完成的模型,还是仅限于存储的视频?能够以书面形式回答这些问题的供应商,拥有的是一套数据治理机制。无法回答的供应商,拥有的只是一个语料库。

在大多数情况下,答案都不会令人安心。但这不是不去提问的理由。恰恰相反,这正是应当在签约之前提出这些问题的理由——那时答案仍能改变决定,而不是等到部署之后,那时答案只能改变责任归属。率先提问的买家,将为整个市场最终必须作答的条件定下基调。

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

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.

引用方式

Grizzle, H. M. (2026, September 27). 语料问题. Novara Consulting Group. https://doi.org/10.5281/zenodo.23003735