The Community and Ethics Shaping the Norwegian Sign Language Corpus

Lindsay Ferrara

NTNU, Norwegian University of Science and Technology
Edvard Bulls veg 1, 7491 Trondheim, Norway
lindsay.n.ferrara@ntnu.no

Abstract

Recently, the Norwegian Sign Language Corpus has been published, and it includes language data from over 100 signers from around Norway. Collecting and building such multimodal signed language corpora have important implications for both research and deaf communities. However, consideration is needed to protect the personal nature of signed language data, while also making a long-term resource that is as accessible as possible to various community, research, and professional stakeholders. In addition, the potential exploitation of corpus resources by commercial and other interests, which are not necessarily aligned with the deaf community itself, must also be deliberated. Here, these seemingly opposing issues and the ethics that surround them are discussed. Current best practices in Open Science (including FAIR and CARE data principles), along with ethical discussions raised by scholars working, for example, in Deaf Studies, are shown to be important in navigating this complex research data landscape.

Keywords: Norwegian Sign Language Corpus, Open Science, data sharing, ethics

1. Introduction to the Norwegian Sign Language Corpus

In 2015, the long-term work of building a linguistic corpus for Norwegian Sign Language began. As is well known, the primary language data of signed language corpora entail video-recordings of the language. These data are also accompanied by (more or less) extensive background information on the signers, annotation files of the recordings, documentation of the corpus, and other supplementary materials. Since many of these data contain personal data and are not anonymized, creating signed language corpora and subsequently publishing them require researchers to engage with a range of legal and ethical concerns. In this paper, an exploration of the specific issues relevant to the creation and publication of the Norwegian Sign Language Corpus is presented. First, a brief overview of the corpus is given. Then the main legal framework along with best practices in Open Science is described. Additional ethical issues relevant in today's research and technology landscape are also raised, before concluding with how the corpus has been published and/or made available to the research community, the Norwegian deaf community, and other stakeholders.

Ten years after the initial work in 2015, the first iteration of the Norwegian Sign Language Corpus has been completed and published in CLARINO, the Norwegian branch of CLARIN.[1] It is comprised of four main datasets, which were collected at different times and under different circumstances. Even so, all the datasets were collected with the intention of creating a long-term resource that could contribute to the documentation of Norwegian Sign Language. These datasets are:

- Data collected as part of Rolf P. Halvorsen's doctoral research (Ferrara and Halvorsen, 2021): collected in 2007, recordings of four signers (two younger and older men and two younger and older women)[2], re-telling of "Frog, Where Are You?" (Mayer, 1969) and personal experience narratives of the 9/11 attacks,
- Pilot project (Ferrara and Bø, 2022): collected in 2015, recordings of seven elderly signers from Trondheim, Oslo, and Bergen (six women and three men, aged between 61-74) having conversations (dyad, triad, and multiparty),
- Depicting Perspective in Norwegian Sign Language (Ferrara and Ringsø, 2021): collected in 2017-2018, recordings of 21 deaf signers (13 women and eight men, aged between 22-57) in conversation (dyad and triad), and
- A Norwegian Sign Language Ecology (Ferrara, 2024a; Ferrara, 2024b): collected in 2019-2025, recordings of 87 signers of all ages (46 women and 40 men, aged 23-91) in elicited and (semi-)spontaneous interactions (dyad and triad).

[1] CLARIN is a European digital infrastructure that supports research based on language resources, https://repo.clarino.uib.no/xmlui/

[2] Exact ages of participants were not archived.

Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages, pages 148–154

Over 100 signers[3] participated in these projects, and they vary by gender, age, education, where they live, and socio-economic background. Some have deaf parents; some have hearing parents. However, they all report being active members of the deaf community and using Norwegian Sign Language in their everyday lives. These signers were recorded engaging in a variety of language-based tasks, for example: informal conversation; personal experience narratives; re-tellings of picture books; interviews focusing on growing up deaf in Norway; as well as discussions on a range of topics relevant to the deaf community.

In total, the Norwegian Sign Language Corpus contains approximately 60 hours of video-recordings, edited into approximately 650 video files. These video files are being annotated in ELAN (Crasborn and Sloetjes, 2008), across approximately 410 ELAN files. The corpus also contains extensive background information on the signers, consent forms, the Norwegian Signbank (Ferrara, 2020), stimuli materials, and other supplementary files that organize and document the corpus (see Ferrara, 2025 for access to some of these non-video materials).

2. The General Data Protection Regulation, Open Science, and research ethics in a digitized world

In the early stages of all signed language corpus projects, a decision must be made on how open and accessible the corpus will be, and what its future uses can entail. As is well known, the primary data of a signed language corpora are video recordings of people using the language—as such, they contain personal data which are legally protected in many places. Relevant to the Norwegian context is the Norwegian Personal Data Act, which, in part, enacts the European Union's General Data Protection Regulation (GDPR). These legal protections work to give individuals more control and agency over their own data and how it is used in today's digitized world.

The enactment of GDPR in Norway in 2018 led to uncertainty in academia regarding (personal) research data. Plans to collect data for the Norwegian Sign Language Corpus was met with high levels of scrutiny by Sikt, The Norwegian Agency for Shared Services in Education and Research, who approve the collection of personal data in research contexts. While this scrutiny was warranted and welcomed, original demands that the data not be stored long-term and restricted to only the purview of the original project went against the goals of collecting a linguistic corpus in the first place.

These initially proposed restrictions were a large contrast to the aims of Open Science, which were also gaining more traction in Norway: "…[O]pen science is defined as an inclusive construct that combines various movements and practices aiming to make multilingual scientific knowledge openly available, accessible and reusable for everyone, to increase scientific collaborations and sharing of information for the benefits of science and society, and to open the processes of scientific knowledge creation, evaluation and communication to societal actors beyond the traditional scientific community» (UNESCO, 2021, p. 7).

One important facet of Open Science is making the data science is based on available to others. Open research data "can be openly used, reused, retained and redistributed by anyone, subject to acknowledgement. Open research data are available in a timely and user-friendly, human- and machine-readable and actionable format, in accordance with principles of good data governance and stewardship...» (UNESCO, 2021, pp. 9-10). Best practices in creating open research data, especially for open signed language data, include the adoption of both FAIR and CARE principles. The FAIR Guiding Principles (Wilkinson et al., 2016; summarized by GO FAIR, no date) recommend that research data (or even non-data objects) be:

- Findable: data can be found by machines and humans, largely through metadata,
- Accessible: information about how the resource can be accessed is provided, with possibly additional authentication and authorization,
- Interoperable: "The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.» (GO FAIR, no date), and
- Reusable: data and metadata need good documentation so that they can be reused in other contexts.

FAIR data supports Open Science goals by facilitating the long-term care of digital research objects that are open and available to 'down-stream' human- and machine-driven activities. They also encourage transparency, reproducibility and re-usability in science. However, while FAIR data should be as open as possible, it should also be as closed as necessary

[3] While most of these signers are deaf, due to the nature of the interaction in some of the recordings, the Corpus also includes two hearing signers.

(UNESCO, 2021, p. 11) to respect, for example, the rights of research participants and their personal data. This tension between protecting personal data while creating research data that was as open as possible guided the initial consideration of how data in the Norwegian Sign Language Corpus was to be collected, processed, and further shared.

Nuance was then added to these seemingly opposing forces (GDPR and data protection on one side, and Open Science on the other) by socio-political realities and ethical considerations. First, as a foreign, hearing researcher it was important to consider how a Norwegian Sign Language Corpus (and subsequent linguistic research) could contribute and align with the goals of the Norwegian deaf community. How would such a long-term resource support and promote this minority language? How could members of the deaf community become involved and hopefully engage in linguistic research using the corpus in the future?

These questions were notably relevant because at the time a new Norwegian Language Act was being drafted and ultimately passed (Språklova 2022). The act, among other things, officially recognizes Norwegian Sign Language as the natural signed language of the deaf community in Norway and stipulates that the Norwegian state has the responsibility to promote and strengthen this language. In response to this law, the government established a committee to survey the status of Norwegian Sign Language in society and to draw up recommendations which would enact the responsibilities outlined in the Language Act (NOU 2023:20). One of the recommendations put forth is the need for more linguistic research and documentation of Norwegian Sign Language. Research work on Norwegian Sign Language is also an important part of the other six recommendations laid out in the report. The Norwegian Sign Language Corpus as a resource along with research resulting can be seen to contribute to this national responsibility of promoting and strengthening Norwegian Sign Language.

Ethical considerations also played a prominent role in decisions around the collection and sharing of the Norwegian Sign Language Corpus. For example, the SLLS Ethics Statement for Sign Language Research[4] raised specific issues relating to research with deaf communities and gives advice about how to respect signers' personal data as well as train, promote, and include deaf people in the research process.

It was also vital to consider how the Corpus might be used (and/or exploited) by technology and AI projects, e.g., machine learning and automatic signed language recognition and translation. There has been a large increase in the number of projects aiming, for example, to develop "assistive technologies" for deaf people, with generally, very little engagement or input from deaf people (De Meulder, 2021; Angelini et al., 2024). Some of these projects have hoped to leverage signed language corpora in their workflows, even with the serious limitations and biases that come with that (De Meulder, 2021; Jantunen et al., 2021; Schembri & Cormier, 2022). While the Norwegian Sign Language Corpus was not created to support developments in machine learning or AI, the high-paced developments across those fields and the ethics surrounding it cannot be downplayed. The European Union of the Deaf's Ethical Framework On AI And Sign Language is relevant here, which demands adherence to *Deaf Digital Law* "a developing legal framework aimed at regulating and safeguarding the equitable and ethical use of digital technologies by and for deaf people» (Venade de Sousa, 2025, p. 13).

3. CARE data principles in the context of deaf signing communities

In addition to the ethical guidelines mentioned in the previous section, I argue here that the CARE data principles can help signed language researchers consider and navigate a complex research data landscape (see also Schulder and Hanke, 2022 on how CARE principles played a role in the creation of the DGS, German Sign Language, Corpus). Originally designed to promote Indigenous Peoples' data sovereignty, the goals of the principles included, "fostering Indigenous self-determination by enhancing Indigenous use of data for Indigenous pursuits» (Carroll et al., 2020, p. 3), and honoring the FAIR data principles. The principles recognize that Indigenous Peoples historically have been subjected to "data inequities and data exploitation" (Carroll et al., 2020, p. 2), which includes the fact that most indigenous data is not held by indigenous people. By increasing control over their own data Indigenous People can reposition themselves from being "subjects of data that perpetuate unequal power distributions to self-determining users of data for development and wellbeing" (Carroll et al., 2020, p. 2). The CARE principles are summarized in Figure 1.

The spirit behind the CARE principles equally apply to deaf signing communities. There should be an aim to foster deaf peoples' self-determination by enhancing deaf people's use of data for deaf pursuits. One can equally recognize that, similar to Indigenous Peoples, deaf people have also been subjected to data inequity and exploitation. We can probably also assume that even today data collected about deaf people are

[4] https://slls.eu/slls-ethics-statement/

still largely controlled by hearing people, perpetuating unequal power distributions.

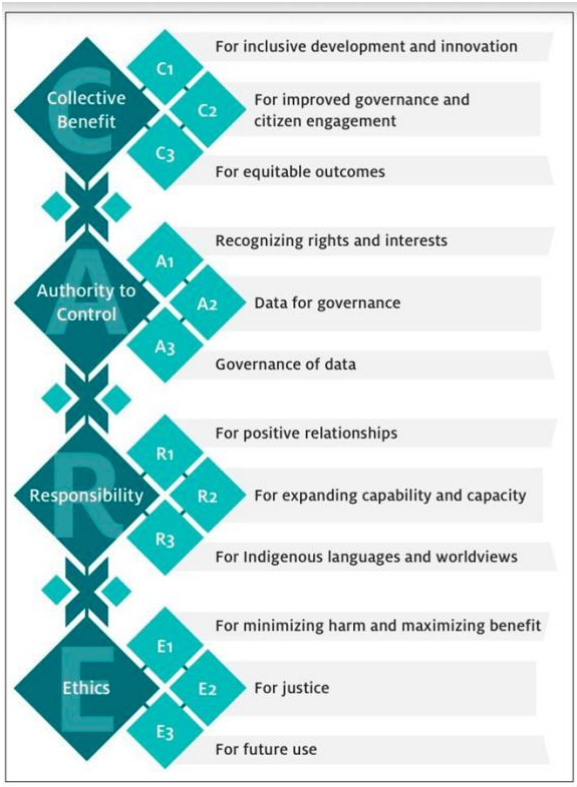

Figure 1: The CARE Principles for Indigenous Data Governance (reproduced from Carroll et al., 2020, p. 5).

The CARE principles stipulate that data should "facilitate the collective benefit" for the community, that communities have an "authority to control and govern" their data including stewardship decisions, that the people working with the data have a responsibility to the communities (e.g., respect, capacity building), and that ethical use of the data should reduce harm and maximize benefit (Carroll et al., 2020, p. 6). The ethical use of data can be further elaborated: "Indigenous Peoples' rights and wellbeing should be the focus across data ecosystems and throughout data lifecycles in order to minimize harm, maximize benefits, promote justice, and allow for future use. Paramount to ethics in data practices is representation and participation of Indigenous Peoples, who must be the ones to assess benefits, harms, and potential future uses based on community values and ethics" (Carroll et al., 2020, p. 6).

Again, it is not a big leap to see how these principles are equally applicable to the world's diverse deaf communities. Researchers working with deaf research data, such as signed language corpora, should facilitate benefit for the deaf community (see also Hochgesang and Palfreyman, 2022 about linguists working with deaf communities). And deaf communities themselves should have authority in the control of such data and how it is used. The ultimate guiding principle should be respect and supporting deaf people's rights and wellbeing.

4. Conclusions: accessing the Norwegian Sign Language Corpus

Before decisions were made about the collection and archival of the Norwegian Sign Language Corpus, careful deliberation over best practices in Open Science along with the relevant legal and ethical obligations, mentioned above, was carried out in collaboration with research and community partners (e.g., discussions were had with international colleagues who have experience building signed language corpora, a focus group of deaf people from the community, the Norwegian Sign Language representative at the Norwegian Language Council). Beginning with the aim that the data be 'as open as possible, and closed as necessary,' the first set of considerations revolved around legal obligations. Ensuring compliance with the EU's GDPR was requisite, so information and consent documents were drafted and revised in collaboration with the Norwegian national service Sikt. The consent documents in particular included specific formulations surrounding the use of personal data within and external to the EU. Also, consent was requested for the long-term archival and use of the data for teaching and research purposes.

In the end, it was decided that the video-recorded data be licensed under two different conditions. First, a Creative Commons license (CC BY-NC-SA 4.0) was applied to data considered non-sensitive. These data include recordings of signers re-telling picture books, a set of personal experience narratives, and public events. In this way, a portion of the corpus data could be used by the (non-academic) deaf community and relevant stakeholders. For example, it was envisioned that the recordings could be used in the future to make teaching materials.

The recordings of conversations, private meetings or events, and discussions were licensed under a more restricted academic license. Among other things this license requires a (research) plan to be submitted to the Norwegian Sign Language Corpus manager detailing how the data will be used.[5] This restricted license was chosen for several reasons. First, it works to protect the signers' and any third persons' privacy (aligning with GDRP). Second, it helps measure

[5] Currently the role of Norwegian Sign Language Corpus manager is held by the author. This role, however, will need to be passed on over time.

engagement with the corpus and keep track of how it is being leveraged by the deaf and academic communities (which may help with future development work). It is important to note that both licenses are non-commercial. More information about the two licenses can be found on the landing pages to the specific datasets in CLARINO.

As a note on timing of the corpus collection: three of the corpus datasets were collected before GDPR came into force. Although participants in these projects did sign consent for their data to be used long-term, in research and teaching, special consideration related to GDPR and the current technological landscape was missing. Therefore, it was decided to reach out to the signers from those previous projects and ask if they would be willing to consent to their data being archived in a (modern) Norwegian Sign Language Corpus, online, in line with GDPR. This process was time-consuming, but worthwhile, because all data now archived in the corpus are made available under the same conditions. The re-consent procedure also gave participants an ability to update their consent (or not) over the use of their data in the current internet age (e.g., data from Ferrara and Halvorsen, 2021, was collected in 2007, well before the age of social media and AI).

Next, also with privacy considerations in mind, it was decided to publish only some aspects of the signer's background information with the video recordings: summarized information about gender, age, where the signer lives, and when they acquired Norwegian Sign Language. These variables are often important for sociolinguistic studies. However, additional details about the signers' background (e.g., language practices in the family and school, work background, etc.) can be made available by contacting the Norwegian Sign Language Corpus manager. Decisions are made case by case.

A final decision was made to not publish the annotation files along with the video data. Although publishing annotation files and video files together are what make a signed language video collection into a corpus (Johnston, 2010), several reasons motivated this decision. First, for a very practical reason, if the annotation files were indeed to be published with the video data, it would mean the video data would not be published for many years. The project team in Norway is very small, with only one primary person working on the data. Thus, it will take many years to complete basic gloss annotation for the corpus videos. In addition, only very little translation work has been carried out, which is also considered requisite initial work (see Johnston, 2024). Out of a perceived need to publish the video data as soon as possible, so that it can be used by the community, it was decided to publish the video data without annotation files. Perhaps, in the future, annotation files can be published, and then versioned, as more work is completed. In the meantime, the annotation files can be shared by contacting the Norwegian Sign Language Corpus manager.

A second reason to not publish the annotation files was made in light of the quickly shifting technological landscape relating to big data, machine learning, and AI. Again, these technologies (in the context of deafness and signed language) often use signed language corpora to train computer models, and this process involves using annotations of the corpus data. Since, for the time being, it seems these projects often are not situated within the deaf community and are not aligned with community values or wishes, it was decided to limit access to the annotation files. As a hearing, foreign researcher, I, as the Norwegian Sign Language Corpus manager, decided it was not my place or role to facilitate such projects. However, there is nothing barring the sharing of the annotation files. Thus, in the future, if a technology project arising from within the deaf community is to be carried out, it is possible to facilitate access to the corpus' annotation files, with the caveat that use aligns with the other license restrictions.

Creating a signed language corpus in the modern age comes with a diverse set of legal and ethical considerations. For the Norwegian Sign Language Corpus, decisions about how to make the corpus accessible to the deaf community was as important as ensuring that the personal data of signers was respected. FAIR and CARE principles helped frame considerations, encouraging a discussion of how the deaf community could be empowered and respected by the project and the resulting corpus. While it was decided that the corpus should be available to as many relevant stakeholders as possible, it was important to consider downstream use of the corpus, especially in regard to engagement and/or exploitation by new technologies. These concerns, along with respect for the personal data collected in the corpus, warranted the implementation of some more conservative protections. It is hoped that now the corpus can be used by the deaf community and researchers to promote, strengthen, and underscore the value and importance of Norwegian Sign Language to the Norway's linguistic landscape.

5. Acknowledgments

I would like to give a heartfelt thanks to all the signers who participated in data collection for the Norwegian Sign Language Corpus, for their time and sharing their language. I would also like to thank all the people who have helped on the different projects as well as the corpus work itself (for a complete list, Ferrara, 2026).

This work has been financially supported by Norwegian Ministry of Culture, NTNU, and the

Norwegian Research Council under project #287067 'Language Use in the Norwegian Deaf Community: Reflections of a Signed Language Ecology.'

6. Bibliographical References

Angelini, R., Spiel, K., and de Meulder, M. (2024). Bridging the gap: Understanding the intersection of deaf and technical perspectives on signing avatars. In A. Way, L. Leeson, and D. Shterionov (Eds.), *Sign language machine translation*. Springer, pp. 291-308. https://doi.org/10.1007/978-3-031-47362-3_12

Carroll, S. R., Garba, I., Figueroa-Rodríguez, O. L., Holbrook, J., Lovett, R., Materechera, S., Parsons, M., Raseroka, K., Rodriguez-Lonebear, D., Rowe, R., Sara, R., Walker, J. D., Anderson, J., and Hudson, M. (2020). The CARE Principles for Indigenous Data Governance. *Data Science Journal, 19*(43):1-12. https://doi.org/10.5334/dsj-2020-043

Crasborn, O., and Sloetjes, H. (2008). Enhanced ELAN functionality for sign language corpora. In Onno Crasborn, Eleni Efthimiou, Thomas Hanke, Ernst D Thoutenhoofd, and Inge Zwitserlood (Eds.), *Proceedings of the LREC2008 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora.* Paris: European Language Resources Association, pp. 39-43. Retrieved from https://www.sign-lang.uni-hamburg.de/lrec/pub/08022.html.

De Meulder, M. (2021). Is "good enough" good enough? Ethical and responsible development of sign language technologies. In Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL), Virtual. Association for Machine Translation in the Americas, pp. 12–22

GO FAIR. (no date). *FAIR principles*. Retrieved 9 February 2026 from https://www.go-fair.org/fair-principles/

Hochgesang, J. A., and Palfreyman, N. (2022). Signed language corpora and the ethics of working with signed language communities. In J. Fenlon and J. A. Hochgesang (Eds.), *Signed language corpora*. Gallaudet University Press, pp. 159-195.

Jantunen, T., Rousi, R., Rainò, P., Turunen, M., Valipoor, M. M., and García, N. (2021). Is there any hope for developing automated translation technology for sign languages? In M. Mämäläinen, N. Partanen, and K. Alnajjar (Eds.), *Multilingual facilitation*. University of Helsinki Library, pp. 61-73. https://doi.org/10.31885/9789515150257

Johnston, T. (2010). From archive to corpus: transcription and annotation in the creation of signed language corpora. *International Journal of Corpus Linguistics*, *15*(1):104-129. https://doi.org/10.1075/ijcl.15.1.05joh

Johnston, T. (2024, November). Auslan corpus annotation guidelines. Manuscript. Sydney: Macquarie University. Retrieved from http://auslan.org.au/about/annotations/

Mayer, M. (1969). *Frog, where are you?* New York: Dial Press.

Schembri, A., and Cormier, K. (2022). Signed language corpora: Future directions. In J. Fenlon and J. A. Hochgesang (Eds.), *Signed language corpora*. Gallaudet University Press, pp. 196-220.

Språklova. (2021). *Lov om språk [The language act]* (LOV-2021-05-21-42). Lovdata. https://lovdata.no/dokument/LTI/lov/2021-05-21-42

NOU 2023:20. (2023). Tegnspråk for livet: Forslag til en helhetlig politikk for norsk tegnspråk [Signed language for life: A proposal for a holistic treatment of Norwegian Sign Language]. Kultur- og likestillingsdepartementet. Retrieved from https://www.regjeringen.no/no/dokumenter/nou-2023-20/id2984187/

Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*, 3:160018. https://doi.org/10.1038/sdata.2016.18

Schulder, M., and Hanke, T. (2022). How to be FAIR when you CARE: The DGS Corpus as a case study of open science resources for minority languages. In Nicoletta Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis (Eds.), *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC2022)*. Marseille, France. European Language Resources Association (ELRA), pp. 164-173. https://aclanthology.org/2022.lrec-1.18/

UNESCO. (2021). *UNESCO Recommendation on Open Science*. https://doi.org/10.54677/MNMH8546

Venade de Sousa, F. (2025). *Sign language in the era of artificial intelligence*. European Union of the Deaf.

7. Language Resource References

Ferrara, L. (2020). *Norwegian Sign Language dataset in Global Signbank.* Nijmegen: Radboud University, Centre for Language Studies. https://signbank.cls.ru.nl/

Ferrara, L. (2024a). *Norwegian Sign Language Corpus – Language Ecology (Retellings and public events)*. Common Language Resources and Technology Infrastructure Norway (CLARINO) Bergen Repository. PID http://hdl.handle.net/11509/155.

Ferrara, L. (2024b). *Norwegian Sign Language Corpus – Language Ecology (Conversations*

and private meetings/events). Common Language Resources and Technology Infrastructure Norway (CLARINO) Bergen Repository. PID http://hdl.handle.net/11509/156.

Ferrara, L. (2025, April 28). Norwegian Sign Language Corpus - Associated files and materials. PID https://doi.org/10.17605/OSF.IO/TYDF4

Ferrara, L. (2026, January). Annotating the Norwegian Sign Language Corpus. Manuscript. Trondheim, Norway: NTNU. Retrieved from https://doi.org/10.17605/OSF.IO/TYDF4

Ferrara, L and Bø, V. (2022). *Norwegian Sign Language Corpus – Pilot Corpus (Conversations)*. Common Language Resources and Technology Infrastructure Norway (CLARINO) Bergen Repository. PID http://hdl.handle.net/11509/147.

Ferrara, L. and Halvorsen, R. P. (2021). *Norwegian Sign Language Corpus - Halvorsen (2012)*. Common Language Resources and Technology Infrastructure Norway (CLARINO) Bergen Repository. PID http://hdl.handle.net/11509/141.

Ferrara, L. and Ringsø, T. (2021). *Norwegian Sign Language Corpus – Depicting Perspective*. Common Language Resources and Technology Infrastructure Norway (CLARINO) Bergen Repository. PID http://hdl.handle.net/11509/144.