

# XSL-HoReCo and GoSt-ParC-Sign: Two New Signed Language - Written Language Parallel Corpora

**Mirella De Sisto**

Tilburg University, the Netherlands  
M.DeSisto@tilburguniversity.edu

**Vincent Vandeghinste**

Instituut voor de Nederlandse Taal  
Leiden, the Netherlands  
and KU Leuven, Belgium  
vincent.vandeghinste@ivdnt.org

**Caro Brosens**

Vlaamse GebarentaalCentrum  
Antwerp, Belgium  
caro.brosens@vgtc.be

**Myriam Vermeerbergen**

KU Leuven, Belgium  
myriam.vermeerbergen@kuleuven.be

**Dimitar Shterionov**

Tilburg University, the Netherlands  
D.Shterionov@tilburguniversity.edu

## Abstract

Developments in language technology targeting signed languages are lagging behind in comparison to the advances related to what is available for so-called spoken languages.<sup>1</sup> This is partly due to the scarcity of good quality signed language data, including good quality parallel corpora of signed and spoken languages. This paper introduces two parallel corpora which aim at reducing the gap between signed and spoken-only language technology: The XSL Hotel Review Corpus (XSL-HoReCo) and the Gold Standard Parallel Corpus of Signed and Spoken Language (GoSt-ParC-Sign). Both corpora are available through the CLARIN infrastructure.

## 1 Introduction

In Europe about half a million people have a sign language as their main or preferred means of communication (Pasikowska-Schnass, 2018). Nevertheless, when talking about language technology, sign language technology is extremely lagging behind in comparison to the tools available for spoken languages (Vandeghinste et al., 2023). One of the reasons is the scarcity of data.<sup>2</sup> This is partially due to the fact that sign languages do not have a widely-used written form, hence collecting written sign language data is not an option (in contrast to what is the case for many spoken languages).

Data collection and data storage also face a number of challenges, such as GDPR<sup>3</sup> restrictions, difficulties in recruiting participants, etc. A lot of short videos are scattered around different platforms and websites, which makes it difficult and time consuming to track them down and get the informed consents of the signers (Vandeghinste et al., forthcoming).

The majority of sign language data comes in the form of videos. To date there is no automatic tool able to annotate or translate sign language videos (Morgan et al., 2022; Vandeghinste et al., 2023), which means that these processes rely on very time-consuming manual work; consequently, the amount of available annotations or translations is scarce.

In addition to that, the quality of the data which are available is often rather problematic (Vandeghinste et al., forthcoming). Most of the sign language datasets readable for machine learning consist of television broadcasts with a spoken language as a source, such as Camgöz et al. (2021) and Koller et al. (2015) which is then interpreted into a sign language by a hearing interpreter. In those cases sign language is the target language of interpreting, which often occurs simultaneously and in real time, and might be influenced by the source language as well as affected by the interpreting process. Most hearing interpreters

<sup>1</sup>We characterize *spoken* language as contrasting with *signed* language, rather than in relation to speech data.

This work is licenced under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>

<sup>2</sup>For a detailed overview of data-related challenges, see De Sisto et al. (2022) and Vandeghinste et al. (forthcoming)

<sup>3</sup>The General Data Protection Regulation is a component of the EU privacy law and human rights law.

do not use a sign language as their main or preferred means of communication (the exception being interpreters who are CODA's – Children of Deaf Adults – and some other specific cases); consequently, they are considered L2 signers.

Additionally, the length of e.g. news broadcasts and the range of specific topics with associated specific lexicon, as well as the speed at which information is disseminated, and the number of names that need to be fingerspelled, all heavily factor into the quality of the result. Interpreters are usually required to take a break every 15-20 minutes when they interpret simultaneously, to keep the quality up and avoid cognitive overload, while news broadcasts are often longer and the pace of information is very high. Different from face to face situations, the interpreter cannot ask the newsreader to repeat themselves or go slower, so in order to keep up with the pace, the interpreters might lean more towards the source text than is ideal.

Within the two projects we present in this paper, we take this into account. Along with the open distribution of these data sets (making them available for the wider research community), the quality of the data (professional translations, involvement of native signers for translation and validation, etc.), and the different (identifiable) domains, they have been collected in a way that suits their use in Machine Learning (ML) applications, and thus have the potential to stimulate the advancements in the field of signed language technology through both high-quality data for training models as well as a gold standard for testing.

After presenting related work in Section 2, we present two recent projects that each aim to address the lack of good quality data by providing parallel data of signed and spoken language data.

- the **XSL Hotel Review Corpus** (XSL-HoReCo) consists of a parallel dataset of hotel reviews in written English (the source language), videos in Sign Language of the Netherlands (Nederlandse Gebarentaal, NGT), Flemish Sign Language (Vlaamse Gebarentaal, VGT), Spanish Sign Language (Lengua de Signos Española, LSE), written Dutch, Spanish and Irish. This dataset is described in Section 3.
- the **Gold Standard Parallel Corpus of Signed and Spoken Language** (GoSt-ParC-Sign), a gold standard dataset of semi-spontaneous Flemish Sign Language (Vlaamse Gebarentaal) (VGT) videos translated into written Dutch. This dataset is described in Section 4.

Section 5 draws conclusions.

## 2 Related work

Although sign languages are low resource languages, there have been some data collection efforts in the past. Kopf et al. (2021) contains a comprehensive list of available corpora for sign languages, but is limited to those cases where sign language is the source. The associated Sign Language Compendium (Kopf et al., 2022)<sup>4</sup> requires as a criterion for inclusion that a corpus must contain (semi-)spontaneous signing, provide transcriptions or translations for at least some of its content and contain at least 10 hours of sign language recordings.

Various sign language datasets have been collected over the years, e.g. CorpusNGT (Crasborn et al., 2008) or DGSKorpus (Prillwitz et al., 2008). However, such datasets are not particularly suited for machine learning or deep learning applications, and require substantial processing prior to building language technology for signed languages (De Sisto et al., 2022; Vandeghinste et al., forthcoming).

For the signed languages addressed in XSL-HoReCo and GoSt-ParC-Sign the following data are available. For NGT existing datasets with authentic signers are the Corpus NGT (Crasborn et al., 2020) and part of the ECHO corpus (Nonhebel et al., 2004). For VGT this is limited to the Corpus VGT (Van Herreweghe et al., 2013). For LSE there is the Corpus de la Lengua de Signos Española (CORLSE)<sup>5</sup> and the small corpus iSignos.<sup>6</sup>

As already mentioned, some sign language datasets that are regularly used for sign language recognition or translation contain *non-authentic* sign language. In these cases we cannot assume that the signers

<sup>4</sup><https://www.sign-lang.uni-hamburg.de/lr/compendium/index.html>

<sup>5</sup><https://corpuslse.es/>

<sup>6</sup><https://http://isignos.uvigo.es/>

belong to the respective sign language community of the language they sign, as most often they are *hearing* sign language interpreters. There is still debate in the sign language technology research community whether the price for using lower quality data can be compensated by the amount of such data, which is much more abundantly available.

Such data for VGT is available in the Content4All corpus (Camgöz et al., 2021). To alleviate this, in the BeCoS data (Vandeghinste et al., 2022) the interpreters are deaf signers re-interpreting hearing signers (which are not on the video), so the resulting sign language, although still being the target language, can be considered authentic. More data has been collected, such as more television broadcasts with sign language interpretation in VGT and videos of the plenary sessions of the Belgian Federal Parliament, with live interpretation into VGT and French Belgian Sign Language (Langue des signes de Belgique francophone; LSFB), but has not yet been processed nor released, partly due to legal constraints (for the broadcasts).

The availability of LSE data is more scattered and not easily gathered. For instance, the corpus created by Porta (2014) contains Spanish texts from different domains which were translated by an interpreter into LSE. However, the video data are not publicly available. Another example of data with limited availability in which LSE is the target language is the material produced by the Fundación CNSE (State Confederation of Deaf),<sup>7</sup> such as an online driving license manual platform.<sup>8</sup> The signed videos are accessible online but the source texts are only visible in the images displaying street signs, hence, not easy to compile nor ML-usable. In most cases, not many metadata are provided concerning the source of the video material. Therefore, it is not immediately possible to evaluate the quality and the authenticity of the signing.

### 3 The XSL Hotel Review Corpus

The XSL Hotel Review Corpus is a multilingual parallel corpus of Sign Language of the Netherlands (Nederlandse Gebarentaal - NGT), Flemish Sign Language (Vlaamse Gebarentaal - VGT), Spanish Sign Language (Lengua de Signos Española, LSE), written English, Dutch, Spanish and Irish.

The focus on a restricted domain ensures recurrence of similar constructions and terms and facilitates the mapping of words or messages to different realisations of signs or signed utterances.

The choice for the domain of hospitality was motivated by the results of co-creation events of the SignON project:<sup>9</sup> during these events, deaf individuals identified the set of circumstances connected to hotels, restaurants, etc., as an appropriate environment in which sign language technology tools would be useful and acceptable. This relates to the concern of some members of the deaf community about the use of these technologies in sensitive or critical situations in which the presence of a human interpreter is preferred.

#### 3.1 Written text

The English source text was taken from the Hotel Reviews dataset publicly available on Kaggle.<sup>10</sup> The original dataset contains a list of 1,000 hotels and their reviews provided by Datafiniti's Business Database. The dataset includes hotel location, name, rating, review data, title, username, and more.

For XSL HoReCo, we only used a selection of the actual hotel reviews. 300 reviews were selected according to the following criteria:

- The review is in English;
- The text is grammatically complete and correct;
- The text does not contain uncommon abbreviations (e.g. *mntns* for 'mountains').

---

<sup>7</sup><https://www.fundacioncnse.org>

<sup>8</sup><https://www.fundacioncnse.org/dgt/>

<sup>9</sup><https://signon-project.eu/>

<sup>10</sup><https://www.kaggle.com/datasets/datafiniti/hotel-reviews>

- In the case in which the review contains incomplete sentences, the removal of these does not affect the meaning of the whole text (an example is provided in table 1).

Original text	Text without incomplete sentence
The Southside Motel and Marina is a diamond in the rough. My room had a comfortable king size bed, nice size fridge, microwave and coffee pot. The room was clean and the staff went out of their way to make sure I always had clean towels, the room and was clean and that I had coffee supplies. The motel owners... More	The Southside Motel and Marina is a diamond in the rough. My room had a comfortable king size bed, nice size fridge, microwave and coffee pot. The room was clean and the staff went out of their way to make sure I always had clean towels, the room and was clean and that I had coffee supplies.

Table 1: Example of reviews with incomplete sentences whose removal does not affect the meaning of the text as a whole.

Within the XSL-HoReCo project, the selected reviews were translated into different languages. The translations from English into Dutch and into Spanish were performed by professional translation companies which used automatic translation (generated by DeepL) followed by in-depth human post-editing. The Irish translation was performed manually by a professional translation company.

XSL-HoReCo consists of 297 hotel reviews, corresponding to 21,464 words in the English source, 22,274 words in Dutch, and 26,469 words in Irish. Only 283 reviews were translated into Spanish,<sup>11</sup> which consists of 20,470 words. A distribution of the length of these reviews is presented in Figure 1 and shows that most reviews have a length between 100 and 350 characters.

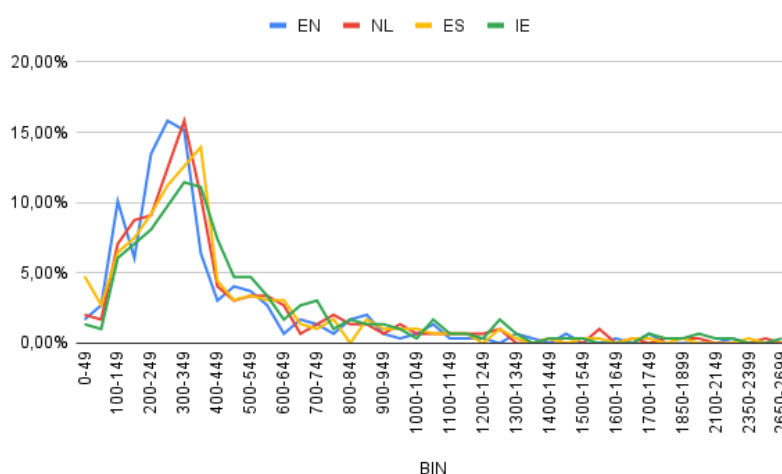


Figure 1: Distribution of the length (in characters) of the different reviews (in bins of 50 characters).

### 3.2 Translation into signed languages

The translations into the three signed languages (i.e. NGT, VGT and LSE) were produced according to the same guidelines (see ‘Translation specifications’ below), concerning the translators, the types of videos produced, and the availability of the data. All translations were made by deaf translators. This reduced as much as possible the interference of the source language. Note that NGT and VGT were translated from the manually post-edited Dutch, while LSE was translated straight from English.

<sup>11</sup>Due to limited budget.

The Dutch text was translated into NGT and VGT by six deaf professional translators each. For the translations into NGT, reviews were shared among translators as shown in table 2. Four translators for the NGT-HoReCo were women. In total, 167 reviews were produced by female signers.

Signer / Translator	No. of videos
P1	50
P2	21
P3	28
P4	49
P5	101
P6	48
Total	297

Table 2: Distribution of videos across NGT-HoReCo signers

For VGT, five were recent graduates from KU Leuven’s training program for deaf translators and interpreters. Translations were divided among translators by assigning to each of them a (close to) equal number of words. Four interpreters were female and two were male.

Translations into LSE were produced by a single translator, due to the very limited availability of deaf professional LSE translators.

Figure 2 shows the distribution of the lengths of the videos for NGT, VGT and LSE, which are mostly between 10 and 60 seconds. Figure 3 shows an example of the videos and texts of the XSL-HoReCo.

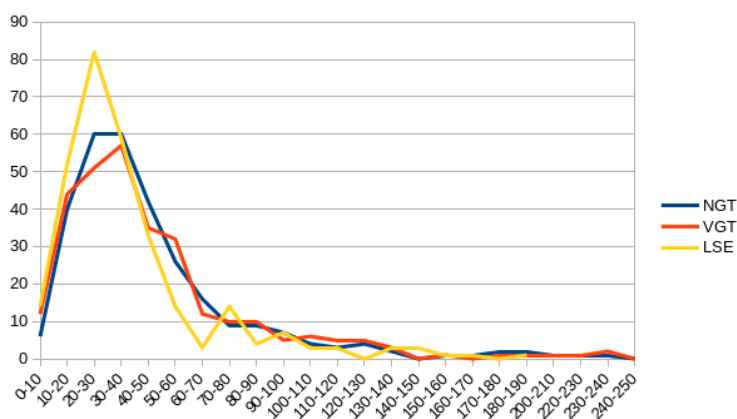


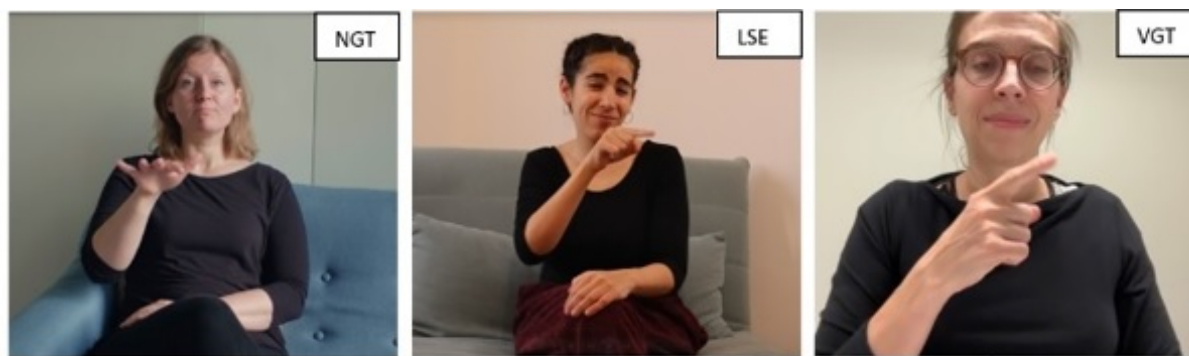
Figure 2: Distribution of lengths of videos (in seconds) in NGT, VGT and LSE

XSL-HoReCo consists of 03:51:45 hours of NGT videos, 03:59:04 hours of VGT videos and 03:09:49 hours of LSE videos, for a total of 11:00:38 hours of recording.

**Translation specifications** Translators were asked to make the recordings in an everyday-life, quiet environment, with a high quality camera. Each video contains one signer translating one review. Each review has been translated once. A future possible expansion of the corpus could include more translations of the same review to better account for inter-signer variation. Nevertheless, given that the corpus focuses on a single domain, i.e. hospitality, a certain recurrence of topics and signs in different possible combinations is already attested; therefore, even if to a limited extent, it allows to account for inter and intra-signer variation.

No time constraint was set for the preparation of a translation before video recording. This was done to ensure the quality of the translation and avoid a “simultaneous-interpretation effect”: during simultaneous interpretation, interpreters are under time-pressure and often need to prioritize efficiency on preserving the complete content of the original message. By having the possibility of preparing the translation

beforehand, XSL-HoReCo translators could make sure that the content of the reviews would be preserved as much as possible during the translation process.



English source text	Automatically translated Dutch text	Postedited Dutch text	Automatically translated Spanish text	Postedited Spanish text	Manually translated Irish text
There's a cool area with shopping and excellent restaurants about half a mile up the road. It's a quick walk but the complimentary y hotel shuttle can zip you over too.	Er is een leuke buurt met winkels en uitstekende restaurants ongeveer een halve mijl verderop. Het is een korte wandeling, maar de gratis hotelshuttle brengt u er ook heen.	Op ongeveer 800 meter afstand is een cool gebied met winkels en uitstekende restaurants. Het is een korte wandeling, maar je bent er ook zo met de gratis hotelshuttle.	Hay una zona interesante con tiendas y excelentes restaurantes a aproximadamente media milla de la carretera. Es una caminata rápida, pero el servicio de transporte de cortesía del hotel también puede acercarlo.	Hay una zona guay con tiendas y excelentes restaurantes a aproximadamente media milla de la carretera. Es una caminata rápida, pero el servicio de transporte gratuito del hotel también puede acercarte.	Tá áit iontach ann ina bhfuil siopaí agus bialanna den scoth thart ar leathmhíle suas an bóthar. Is siúlóid ghairid é ach is féidir leis dul ansin ar an tointeáil óstáin saor in aisce freisin.

Figure 3: Example from NGT, LSE and VGT-HoReCo

### 3.3 Availability

To ensure the availability of the data for future research, all translators signed an informed consent form agreeing with the data being publicly available under the CC-BY NC licence. The accessibility and format of the XSL-HoReCo project makes it easily expandable with more parallel languages and/or additional annotations.

The NGT side of the HoReCo is available through the European Language Grid at <https://live.european-language-grid.eu/catalogue/corpus/21535> and through CLARIN at <http://hdl.handle.net/10032/tm-a2-x7>.

The VGT side of the HoReCo is available through the European Language Grid at <https://live.european-languagegrid.eu/catalogue/corpus/23007>, and through CLARIN at <http://hdl.handle.net/10032/tm-a2-y3>.

The LSE side of the HoReCo is available through the European Language Grid at <https://live.european-language-grid.eu/catalogue/corpus/23263> and is available through CLARIN at <http://hdl.handle.net/10032/tm-a2-x6>.

## 4 GoSt-ParC-Sign

The Gold Standard Parallel Corpus of signed and spoken language focuses on spontaneous and semi-spontaneous VGT and its translation into written Dutch.

The GoSt-ParC-Sign project was developed in three phases: data gathering, manual translation, and quality control. All phases were coordinated and overseen by the Vlaamse GebarentaalCentrum (Flemish Sign Language Centre).

### 4.1 Phase 1: Data collection

During the first phase, roughly ten hours of publicly available semi-spontaneous VGT videos were initially identified. All VGT material contained in this corpus was produced by deaf authentic signers for a signing audience.<sup>12</sup> Therefore, the quality of the signing is as close as it could possibly be to real life signing. Written consent was gathered from the authors and signers of these videos to ensure that we would be allowed to redistribute the material.

The final content of the corpus is presented in Table 3 and amounts to just about 10 hours of footage.

Corpus Part	Duration
Spontaneous conversation from the VGT corpus	3:11:05
Talkshow "Dagelijks Doof"	2:24:24
Vlog regarding typical language use in VGT	1:15:24
Game show "wie wordt miljonair"	1:07:00
Various research rapports professionally translated into VGT	1:46:06
Opinion pieces in VGT	0:13:18
Total	9:57:07

Table 3: Description of the sources of the differnt parts in the GostParc-Sign corpus

The footage contains 43 different signers of different ages and different regions, as presented in Table 4. Age groups are presented in Table 5. This information is relevant for future work on this corpus because much variation and differences are attested across different regions and age groups. Since the corpus only contains already existing data, we had no real control over the distribution of these sociolinguistic factors.

Region	Men	Women
West-Flanders	4	5
East-Flanders	8	7
Flemish-Brabant	1	0
Antwerp	8	4
Limburg	1	3
Total	22	21

Table 4: GostParc signers: origin and gender

The data in the two languages are aligned at the sentence (or message) level, since there is no one-to-one correspondence between VGT signs and Dutch words.

### 4.2 Phase 2. Manual translation

The second phase focused on the manual translation task from VGT into Dutch text. Translations were performed by mixed teams of deaf and hearing translators, in total four deaf and six hearing translators were involved. Having a mixed team had a double purpose:

- i. ensuring that the original meaning of the signed message was preserved through the deaf translator;

<sup>12</sup>For sign languages, it is problematic to talk about native language, since most deaf children are born in hearing families and get exposed to a sign language only later on in life. Consequently, we prefer the term 'authentic', indicating that the individual uses a sign language as their main and preferred language.

Age group	Men	Women
12-18	3	7
19-25	3	4
26-35	7	5
36-50	2	4
51-70	6	2

Table 5: Age groups (at the time of recording)

ii. providing a good quality Dutch text, through the native Dutch translator.

Translations were organised in ELAN (Sloetjes & Wittenburg, 2008), which allows to synchronise multiple annotation tiers with the video timeline. A ‘Translation’ tier was created for each of the participants in the video to contain the written Dutch translation in each ELAN Annotation Format (EAF) file of each video (an example of the format is provided in Figure 4). The image shows one tier for each of the four participants in the talkshow, this way even overlapping utterances could be correctly captured (as seen in the image). Having files in EAF can serve for linguistic research; in addition, this format can be easily adjusted into an ML-suited format with the framework proposed in De Sisto et al. (2022).

Our initial estimation was that 133 hours of translation work would lead to approximately 9–10 hours of videos being translated. This estimation was based on a consultation with professional signed language to spoken language translators, according to which we concluded that 15 minutes of translation work would correspond roughly to one minute of video translation. Unfortunately, during the translation phase we realised that, given the breadth of the topics covered in the video and the spontaneity of the signing, more time was needed for translating them. In total between 180 and 220 hours of translations lead to 8 hours of videos being translated. Consequently, in order to reach the target of having 10 hours of videos, we opted for including in the corpus additional footage publicly available with subtitles.

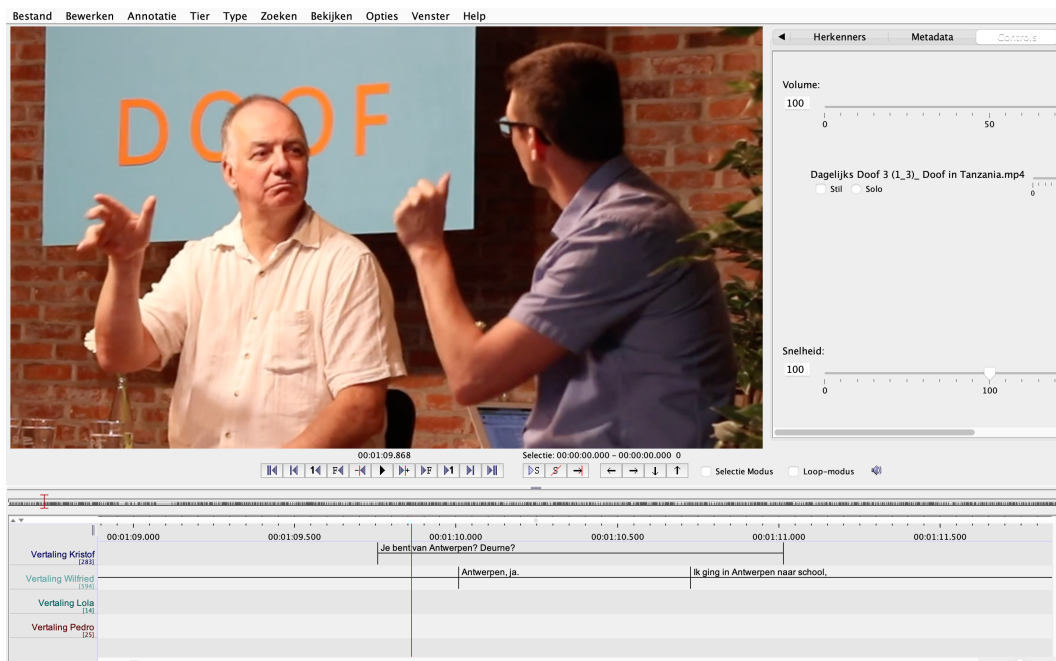


Figure 4: Example of GoSt-ParC-Sign’s data format



### 4.3 Phase 3: Quality control

In the third phase the quality control of the translations was performed by a professional editor who was not part of the initial translation team, to ensure that the produced translations were correctly reporting the original message of the VGT videos and that the Dutch texts were of high quality.

This corpus is made available under CC BY license, at the Instituut voor de Nederlandse Taal (INT) at <http://hdl.handle.net/10032/tm-a2-x9> and will soon be made available on the European Language Grid.

## 5 Conclusion

In this paper we have introduced two signed language data collection projects which aim at supporting advances in more inclusive language technology which also targets signed languages. The XSL-HoReCo project led to the creation of a multilingual parallel corpus of NGT, VGT, LSE, English (source text), written Spanish, Irish and Dutch. The very recently concluded GoSt-ParC-Sign project produced a parallel corpus of authentic VGT videos and a translation into written Dutch. The creation of similar parallel data is fundamental for supporting research and developments into fields such as signed language translation, recognition and processing.

In addition, another important outcome of these data collection project is the lesson learned throughout the process and from the challenges encountered, which can be useful for future high quality signed language data collection projects:

- Guidelines for recording a video need to be extremely clear and specific; potential vagueness might lead to differences in the quality, style and type of the recording.
- It is quite difficult, if not impossible, to have an exact estimation of the ratio of translation time needed per hour of signed language videos. Many factors are at play, which affect the translation process, such as topics discussed, spontaneity of the signing, monologue vs. group conversation, potential peculiarities of individual signing styles, etc. In the GoSt-ParC-Sign project, our initial estimation turned out to be dramatically lower than the actual time needed by translators.
- The translation of spontaneous signing can be particularly challenging. Just as with spontaneous speech, unplanned signing might contain unnecessary repetitions, unclear articulations, false starts; in some circumstances, identifying what is being signed can be challenging even for an expert authentic user of a signed language, independently from whether the user is deaf, hard of hearing or hearing.
- Signers remain at all time the owners of the data they produce.

## Acknowledgements

The initial NGT-HoReCo project (containing NGT and English and Dutch) has been funded by the SRIA contribution Projects of the European Language Equality 2 project.<sup>13</sup> Extensions to VGT, LSE, Irish and Spanish were funded by the SignON project which has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 101017255.

The GoSt-ParC-Sign project has been awarded the EAMT Sponsorship of Activities 2022 and is partially funded by the SignON project.

## References

Camgöz, N. C., Saunders, B., Rochette, G., Giovanelli, M., Inches, G., Nachtrab-Ribback, R., & Bowden, R. (2021). Content4all open research sign language translation datasets. *CoRR*, *abs/2105.02351*. <https://arxiv.org/abs/2105.02351>

<sup>13</sup><https://european-language-equality.eu/>

- Crasborn, O., Zwitserlood, I., & Ros, J. (2008). Het Corpus NGT. Een digitaal open access corpus van filmpjes en annotaties van de Nederlandse Gebarentaal. Nijmegen: Centre for Language Studies, Radboud University. <https://www.corpusngt.nl/>
- Crasborn, O., Zwitserlood, I., Van der Kooij, E., & Bank, R. (2020). *Annotation conventions for the Corpus NGT* (tech. rep.). Radboud University Nijmegen, Centre for Language Studies and Department of Linguistics.
- De Sisto, M., Vandeghinste, V., Egea Gómez, S., De Coster, M., Shterionov, D., & Saggion, H. (2022). Challenges with sign language datasets for sign language recognition and translation. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2478–2487. <https://aclanthology.org/2022.lrec-1.264>
- Koller, O., Forster, J., & Ney, H. (2015). Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, 141, 108–125.
- Kopf, M., Schulder, M., & Hanke, T. (2021). *Overview of datasets for the sign languages of Europe* (Project deliverable No. D6.1). EASIER Consortium. <https://doi.org/10.25592/UHHFDM.9560>
- Kopf, M., Schulder, M., & Hanke, T. (2022). The Sign Language Dataset Compendium: Creating an overview of digital linguistic resources. In E. Efthimiou, S.-E. Fotinea, T. Hanke, J. A. Hochgesang, J. Kristoffersen, J. Mesch, & M. Schulder (Eds.), *13th international conference on language resources and evaluation (LREC 2022). Proceedings of the 10th workshop on the representation and processing of sign languages: Multilingual sign language resources* (pp. 102–109). European Language Resources Association (ELRA). <https://www.sign-lang.uni-hamburg.de/lrec/pub/22025.pdf>
- Morgan, H. E., Crasborn, O., Kopf, M., Schulder, M., & Hanke, T. (2022). Facilitating the spread of new sign language technologies across Europe. In E. Efthimiou, S.-E. Fotinea, T. Hanke, J. A. Hochgesang, J. Kristoffersen, J. Mesch, & M. Schulder (Eds.), *13th international conference on language resources and evaluation (LREC 2022). Proceedings of the 10th workshop on the representation and processing of sign languages: Multilingual sign language resources* (pp. 144–147). European Language Resources Association (ELRA). <https://www.sign-lang.uni-hamburg.de/lrec/pub/22026.pdf>
- Nonhebel, A., Crasborn, O., & van der Kooij, E. (2004). *Sign language transcription conventions for the echo project. annotation convention. version 9* (tech. rep.). University of Nijmegen. <http://hdl.handle.net/2066/57889>
- Pasikowska-Schnass, M. (2018). *Sign languages in the EU* (tech. rep.). European Parliamentary Research Service. [http://www.europarl.europa.eu/RegData/etudes/ATAG/2018/625196/EPRS\\_ATA\(2018\)625196\\_EN.pdf](http://www.europarl.europa.eu/RegData/etudes/ATAG/2018/625196/EPRS_ATA(2018)625196_EN.pdf)
- Porta, J. (2014). *Towards a rule-based Spanish to Spanish sign language translation: from written forms to phonological representations* [Doctoral dissertation, Universidad Autónoma de Madrid. Departamento de Tecnología Electrónica y de las Comunicaciones].
- Prillwitz, S., Hanke, T., König, S., Konrad, R., Langer, G., & Schwarz, A. (2008). DGS corpus project—development of a corpus based electronic dictionary German Sign Language/German. *3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora*, 159.
- Sloetjes, H., & Wittenburg, P. (2008). Annotation by category: ELAN and ISO DCR. *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. [http://www.lrec-conf.org/proceedings/lrec2008/pdf/208\\_paper.pdf](http://www.lrec-conf.org/proceedings/lrec2008/pdf/208_paper.pdf)
- Van Herreweghe, M., Vermeerbergen, M., Demey, E., De Durpel, H., Nyffels, H., & Verstraete, S. (2013). *Corpus Vlaamse Gebarentaal: Annotatierichtlijnen* (tech. rep.). Universiteit Gent and KU Leuven.
- Vandeghinste, V., De Sisto, M., Egea Gómez, S., & De Coster, M. (forthcoming). Challenges with sign language datasets. In A. Way, L. Leeson, & D. Shterionov (Eds.), *Sign language machine translation*. Springer.

- Vandeghinste, V., De Sisto, M., Kopf, M., Schulder, M., Brosens, C., Soetemans, L., Omardeen, R., Picron, F., Van Landuyt, D., Murtagh, I., Avramidis, E., & De Coster, M. (2023). *Report on Europe's Sign Languages* (tech. rep.). European Language Equality D1.40. <https://zenodo.org/records/8047005Domain>
- Vandeghinste, V., Van Dyck, B., De Coster, M., Goddefroy, M., & Dambre, J. (2022). Becos corpus: Belgian covid-19 sign language corpus. a corpus for training sign language recognition and translation. *Computational Linguistics in the Netherlands Journal*, 12, 7–17. <https://clinjournal.org/clinj/article/view/144>