NGT-HoReCo and GoSt-ParC-Sign: Two new Sign Language - Spoken Language parallel corpora

Mirella De Sisto, Dimitar Shterionov
Tilburg University, the Netherlands
M.DeSisto/D.Shterionov
@tilburguniversity.edu

Lien Soetemans
KULeuven, Belgium
lien.soetemans@kuleuven.be

Vincent Vandeghinste
Dutch Language Institute
Leiden, the Netherlands
vincent.vandeghinste@ivdnt.org

Caro Brosens
Vlaams Gebarentaalcentrum
Antwerp, Belgium
caro.brosens@vgtc.be

Abstract

Language technology targeting both signed and spoken languages is extremely limited. This is partly due to the scarce availability of good quality signed language data and of signed and spoken parallel corpora. In this paper we introduce two projects which aim at reducing the gap between spoken-only language technology and more inclusive language technology for both signed and spoken languages by creating two parallel corpora with a sign language on one side and a spoken language on the other: the Dutch - Sign Language of the Netherlands Hotel Review Corpus (NGT-HoReCo) and the Gold Standard Parallel Corpus of Signed and Spoken Language (GoSt-ParC-Sign). Both corpora are or will be made available through the CLARIN infrastructure.

1 Introduction

In Europe about half a million people have a Sign Language (SL) as their main or preferred means of communication (Pasikowska-Schnass, 2018). Nevertheless, when talking about language technology, SL 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 (for a detailed overview of data-related challenges, see De Sisto et al., 2022; Vandeghinste et al., forthcoming); this is partially due to the fact that SLs do not have a widely-used written form used by Deaf communities, hence spontaneous written data is not an option (as it is the case for many spoken languages). Data collection and data storage also face a number of challenges, such as GDPR restrictions, difficulties in recruiting participants, etc.

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

In addition to that, often the quality of the data available is rather problematic (Vandeghinste et al., forthcoming). Most of the ML-readable SL datasets are news broadcast original spoken language interpreted by hearing interpreters, which is rather problematic in terms of the quality of the data: firstly, in those cases SL is the target language of interpreting which often occurs simultaneously, hence, is both influenced by the source language as well as affected by the interpreting process; secondly, most hearing interpreters do not use a SL as their main or preferred means of communication (the exception being interpreters being CODA's – Children of Deaf Adults – and some other specific cases); consequently, they can be considered L2 signers.

In this paper we present two recent projects which address the lack of good quality data by providing two parallel corpora of signed and spoken language data: the Dutch - Sign Language of the Netherlands Hotel Review Corpus (NGT-HoReCo) which consists of a parallel dataset of hotel reviews in written English, written Dutch (translations of the original English by a professional translation service in Dutch), and Sign Language of the Netherlands (Nederlandse Gebarentaal, NGT) videos; the Gold Standard Parallel Corpus of Signed and Spoken Language (GoSt-ParC-Sign), a golden standard dataset of

semi-spontaneous Flemish Sign Language (Vlaamse Gebarentaal) (VGT) videos translated into written Dutch. Such datasets, that include SL data produced by native signers and have been collected in a way that suits their use in ML applications, have the potential to stimulate the advancements in the field of SL technology through both high-quality data for training models as well as a gold standard for testing.

2 NGT-HoReCo

The NGT-HoReCo project took place between January and March 2023. The corpus consists of hotel reviews in written English, translated into written Dutch and into NGT videos. The Dutch text was produced by translating Booking.com hotel reviews from English to Dutch. These reviews are publicly available on Kaggle. [1] The English-Dutch translations were produced by a professional translation company which used automatic translation (generated by DeepL) following an in-depth human post-editing. Dutch to NGT translations were performed by six deaf professional translators. Relying on deaf and not hearing interpreters we ensured that (i) there is as little as possible interference of the source language (Dutch) and (ii) the signing is authentic, i.e. produced by a native (L1) signer. The corpus consists in 283 reviews: 19,950 words in the English source, 21,825 words, on the Dutch text side, and 213.18 minutes on the NGT video side. The advantage of providing data focusing on a single domain, i.e. hospitality, allows to have recurrent topics and signs in different possible combinations and to account, to a certain extent, for inter and intra signer variation.

Figure 1 shows an example of the parallel texts and video. One folder contains all videos. An excel file contains the original English text, a Dutch translation obtained with machine translation, the Dutch translation produced by the translation company; the last two columns contain the video identifier and the signer identifier, respectively.

The hotel was beautiful and the staff was awesome. One of the best beaches in mexico	Het hotel was prachtig en het personeel was geweldig. Een van de beste stranden in Mexico	Het hotel was prachtig en het personeel was geweldig. Een van de beste stranden in Mexico.	NGT-HoReCo_89	P3

Figure 1: Example from NGT-HoReCo

The corpus is available at http://hdl.handle.net/10032/tm-a2-w2 under CC BY-NC license. A CMDI record has been made which should be harvested by the CLARIN Virtual Language Observatory to ensure findability of the corpus. The corpus is also available through the European Language Grid at https://live.european-language-grid.eu/catalogue/corpus/21535.

[1] https://www.kaggle.com/datasets/datafiniti/hotel-reviews

3 GoSt-ParC-Sign

The GoSt-ParC-Sign project started in February 2023 and will be ongoing until January 2024. The corpus will contain videos of spontaneous and semi-spontaneous VGT produced by deaf individuals who use VGT as their main or preferred means of communication for deaf or signing audience. The project consists of three phases.

In the first phase, we identified roughly ten hours of publicly available (semi-)spontaneous VGT videos. These videos cover different topics and genres, such as five hours of free conversation, one and a half hour of panel discussion about linguistic change in the community, over two hours of a deaf-lead talk, a game show to celebrate 15 years of recognition for VGT, and 45 minutes of semi-spontaneous vlogs about typical language uses in VGT. Currently, informed consents for the public availability of the videos are being collected from the video owners. In addition, we recruited a mixed team of deaf and hearing professional VGT translators; having both deaf and hearing translators makes sure that the content of the source is preserved, and ensures good quality of the target translation.

The second phase will focus on translating the VGT videos into written Dutch text. Translations will be organised in ELAN (Sloetjes & Wittenburg, 2008), which allows multiple annotation tiers synchronised with the video timeline. A 'Translation' tier will contain the written Dutch translation in each ELAN Annotation Format (EAF) file of each video (an example of the format is provided in Figure 2). 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).

In the third phase, the coordinators of the translation team together with members of the VGT Deaf community will perform quality control of the translations produced.

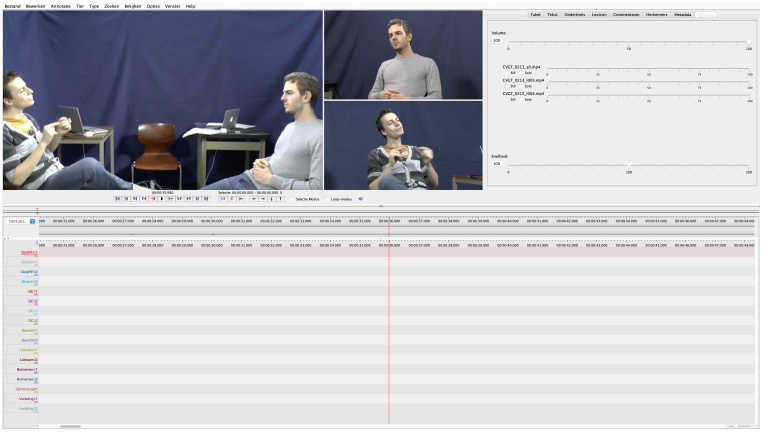

Figure 2: Example of GoSt-ParC-Sign's data format

The corpus will be made publicly available, under CC BY license, at the Instituut voor de Nederlandse Taal (INT), which ensures long-term availability. The metadata will also be published in CMDI formats for harvesting by the CLARIN infrastructure.

4 Use-case

Various SL 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 SLs (De Sisto et al., 2022; Vandeghinste et al., forthcoming). Within these two projects 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 ML applications, and thus have the potential to stimulate the advancements in the field of SL technology through both high-quality data for training models as well as a gold standard

for testing. For example, we have already initiated the further development of the NGT-HoReCo corpus to cover VGT, different type of annotations, pose estimates, etc., to facilitate ML and DL applications. Within the GoSt-ParC-Sign we will use ELAN, following standards to allow for the straightforward use of the data by linguists (familiar with ELAN and the EAF) as well as DL/ML practitioners using tools such as De Sisto et al., 2022.

5 Conclusion

In this paper we have introduced two SL data collection projects which aim at supporting advances in more inclusive language technology which also targets SLs. The very recently concluded NGT-HoReCo project led to the creation of a Dutch - NGT parallel corpus which contains 283 hotel reviews in written English, Dutch and NGT videos. The GoSt-ParC-Sign project is still ongoing and aims at creating a parallel corpus of authentic VGT videos and a translation into written Dutch. The creation of similar parallalel data is fundamental for supporting research and developments into fields such as SL translation, recognition and processing.

Acknwoledgements

The NGT-HoReCo project has been funded by the SRIA Contribution Projects of the ELE 2 project. The GoSt-ParC-Sign project has been awarded the EAMT Sponsorship of Activities 2022 and is partially funded by the SignON project, funded by the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 101017255.

References

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/

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

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.), *Proceedings of the LREC2022 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

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

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

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.

Vandeghinste, V., Sisto, M. D., Gómez, S. E., & Coster, M. D. (forthcoming). *Challenges with sign language datasets*.