GoSt-ParC-Sign
Gold Standard Parallel Corpus of Sign and spoken language

Mirella De Sisto[*], **Vincent Vandeghinste**[†], **Lien Soetemans**[‡], **Caro Brosens**[§], **Dimitar Shterionov**[*]
[*]Tilburg University, [†]Instituut voor de Nederlandse Taal, [‡]KU Leuven, [§]Vlaams Gebarentaalcentrum
m.desisto@tilburguniversity.edu, vincent@ccl.kuleuven.be,
lien.soetemans@kuleuven.be, caro.brosens@vgtc.be,
d.shterionov@tilburguniversity.edu

1 Introduction

In the last decade, there has been an increasing interest in extending MT from only focusing on Spoken Languages (SpLs) to also targeting Sign Languages (SLs); nevertheless, the advances of this field are still limited, and this is due to a number of reasons (e.g. challenges related to data availability, lack of notation conventions, etc.).

Besides the technological gap between SpLMT and SLMT, a severe difference lies in the availability of high-quality (training) data. SpLMT can count on open and free datasets, such as Europarl (Koehn, 2005) and OPUS (Tiedemann and Nygaard, 2004), and on several MT platforms which allow training on specific datasets.[1] The availability of sufficient amounts of high-quality (training) data drives the MT performance up. Furthermore, well-designed test sets allow to adequately assess quality and fairly compare MT systems.

For SLs, instead, training data is scarce and scattered. Parallel datasets, with one side in a SL and the other in a SpL, are extremely limited. In addition, most of the available datasets consist in broadcasts with subtitles/autocues as a written form of a SpL as the source and interpretation into a SL as the target (Camgoz et al., 2018); this leads to various concerns related to their quality: SL as the result of interpretation or translation is heavily influenced by the source language[2] as well as by the interpreting process; in addition, even though in some cases hearing interpreters are CODA's (children of deaf adults), most often the interpretation is made by a hearing interpreter for whom the SL is the L2.

In some cases, corpora with SL as source are available, such as the Corpus Vlaamse Gebarentaal[3] (VGT) (Van Herreweghe et al., 2015) (Corpus of Flemish Sign Language); nevertheless, as annotation of the data is ongoing, the translations available are too insufficient for quality (automatic) SL translation (SLT). Additionally, as the data contain videos of the signer's faces, strict GDPR rules apply, and signed informed consent forms are required from each of the signers.

The SignON project[4] aims to build SLT engines and hence gathers available SL data; throughout this process, we faced a number of issues,[5] which led us to identify the need for a gold standard parallel corpus of SL - SpL. The collection, organisation and (public) release of such a corpus, will provide a common ground for advancing the field of SLT.

2 Gost-Parc-Sign

The goal of this project is to create a gold standard parallel corpus of authentic VGT as source and a translation into written Dutch as target language. This 12-month project, running between February 2023 and January 2024, consists of three phases: (1) Collection of existing source SL videos in VGT and of informed consent forms from their signers.[6] (2) Manual translation of the SL into

[1] See, for instance, Nematus (`https://github.com/EdinburghNLP/nematus`), OpenNMT (`https://opennmt.net/`), MarianMT (`https://marian-nmt.github.io/`),

[2] This phenomenon is referred to as *translationese* (Graham et al., 2020)

[3] `https://www.corpusvgt.be/`

[4] `https://signon-project.eu/`

[5] for an overview of data-related challenges of SLMT see (De Sisto et al., 2022)

[6] Informed consent for the voice over will not be needed, since audio will not be included in our corpus.

written Dutch, performed by a mixed team of deaf and hearing professional VGT translators; this will optimize the translation process, preserve the content of the original message, and ensure good quality of the Dutch text. This phase will consist of 133 hours of translation work,[7] resulting in approximately at least 9–10 hours of video being translated.[8] Translations will be created in ELAN (Sloetjes and Wittenburg, 2008). Translations will be arranged into a "Translation" tier in the ELAN Annotation Format (EAF) file of each corresponding video. Since there is no sign-to-word correspondence between VGT and Dutch, alignment is at the sentence or message level. (3) Quality control by members of the Flemish deaf community and L1 Dutch language users, which will ensure that the translations made convey the same message as the original videos. All phases will be overseen by the Vlaams GebarenTaalCentrum (VGTC) and KU Leuven, both members of SignON, in order to ensure data and translation quality. The final corpus will be made publicly available (with a Creative Commons BY licence) through the CLARIN infrastructure at the Instituut voor de Nederlandse Taal (INT), and through the European Language Grid.

3 Current and future steps

In this initial phase of GoSt-ParC-Sign approximately 10 hours of authentic VGT videos to be translated into written Dutch have been identified. The videos cover different topics and genres: 5 hours of free conversation, a 1,5 hour panel discussion about linguistic change in the community, over 2 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. They all constitute content originally produced for a signing audience. VGTC has recruited translators and we are currently collecting signed informed consents from the video's owners. After phase 1, the translation phase will start; the quality control, i.e. phase 3, will follow between August and December 2023. In the final month of the project we will prepare and release all the data and documentation.

Acknowledgements

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

References

Camgoz, Necati Cihan, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural sign language translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, USA, 18 – 22 June. IEEE.

De Sisto, Mirella, Vincent Vandeghinste, Santiago Egea Gómez, Mathieu De Coster, Dimitar Shterionov, and Horacio Saggion. 2022. Challenges with sign language datasets for sign language recognition and translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2478–2487, Marseille, France, June. European Language Resources Association.

Graham, Yvette, Barry Haddow, and Philipp Koehn. 2020. Statistical Power and Translationese in Machine Translation Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 72–81, Online. Association for Computational Linguistics.

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand, September 13–15.

Sloetjes, Han and Peter Wittenburg. 2008. Annotation by category: ELAN and ISO DCR. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May. European Language Resources Association (ELRA).

Tiedemann, Jörg and Lars Nygaard. 2004. The OPUS corpus - parallel and free: `http://logos.uio.no/opus`. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, pages 1183–1186, Lisbon, Portugal, May. European Language Resources Association (ELRA).

Van Herreweghe, Mieke, Myriam Vermeerbergen, Eline Demey, Hannes De Durpel, Hilde Nyffels, and Sam Verstraete. 2015. Het Corpus VGT. Een digitaal open access corpus van video's and annotaties van Vlaamse Gebarentaal, ontwikkeld aan de Universiteit Gent ism KU Leuven. https://www.corpusvgt.ugent.be/.

[7] This amount was calculated based on the funding available and translators' average hourly rate (60 euro).

[8] This estimate was made by consulting professional SL to SpL translators: 15 minutes of translation work correspond roughly to one minute of video translation. In terms of resulting text, we could estimate, based on a recently concluded corpus project, that the translation of these videos into written Dutch might correspond approximately to 50.000 words.