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

Mirella De Sisto*, Vincent Vandeghinste†‡, Caro Brosens§,
Lien Soetemans‡, 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

June 2023

1 Introduction

Multimodal Machine Translation (MT) which targets both signed and spoken languages is still in its infancy for a number of reasons, related to both the linguistic properties of signed languages (SLs) —for which spoken-language-based MT pipelines are not immediately suited[1] —, as well as to the data (e.g. lack of widely accepted writing system, lack of notation conventions, and of data formatting standards[2]). One major challenge lies in the severe lack of high-quality (training) data for SLs.

MT focusing on spoken languages can count on open and free datasets (e.g. Europarl (Koehn, 2005), OPUS (Tiedemann and Nygaard, 2004) and on various MT platforms that can be used for training on specific datasets (e.g. Nematus,[3] MarianMT[4]). The availability of high-quality data allows to obtain high performance of MT models, and to compare and assess quality of MT systems.

MT also targeting SLs, instead, cannot count on a similar amount of available resources. Parallel datasets for SLs, with one side in a SL and the other in a spoken language, are extremely limited. In addition, most of the available

[1] For details see De Sisto et al. (2021).
[2] for an overview of data-related challenges of SLMT see De Sisto et al. (2022)
[3] `LP/nematus`
[4] `https://marian-nmt.github.io/`

datasets have a SL as the target language obtained with simultaneous interpretation (Camgoz et al., 2018), which raises a number of concerns in terms of authenticity of the signing. Simultaneous interpretation is an extremely time-constrained activity which requires interpreters to make quick decisions for the most efficient solution (which might not always be the most accurate translation). Moreover, given that the SL is almost always the target language in these datasets, it is generally heavily influenced by the source language, leading to what is referred to as *translationese* (Graham et al., 2019, 2020). Finally, another aspect which needs to be considered is that very often these simulaneous interpretations are performed by hearing interpreters for whom the SL is an L2 (made exception for those interpreters who are CODA's —Children of Deaf Adults —and other specific cases). Therefore, there is a high risk of developing MT systems which are trained on L2 users data only.

In some cases, corpora with SL as source are available, such as the Corpus Vlaamse Gebarentaal[5] (VGT) (Van Herreweghe et al., 2015) (Corpus of Flemish Sign Language); however, since often translation and annotation of the data are still ongoing, what is currently available for MT is not sufficient for high quality results. In addition, the data contain videos of the signer's, hence their sharing and usage are restricted by GDPR rules, and signed informed consent forms are required from each of the signers.

Within the SignON project,[6] which aims to build SL MT engines by using available SL data, we have been facing numerous data-related challenges. With the Gold Standard Parallel Corpus of Signed and Spoken language (GoSt-ParC-Sign) we contribute to reducing the lack of good quality data by releasing a publicly available parallel corpus in which semi-spontaneous SL is the source language.

2 Gold Standard Parallel Corpus of Signed and Spoken language

The Gost-ParC-Sign is a multimodal corpus of VGT as source and a translation into written Dutch as target language.

All VGT material included in this corpus consists of already existing VGT videos which were produced by authentic VGT signers for a signing audience. Therefore, they are as close as they could possibly be to spontaneous real life VGT signing.

The corpus contains 10 hours of videos gathered from different sources. Table 1 provides an overview of source and amount of footage.

The signing in VGT is produced by 22 male and 21 female authentic users from different regions of Flanders, as shown in table 2. Signers are from different age groups, as displayed in table 3.

The data in the two languages are aligned at the sentence (or message level),

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

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

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 1: Main source of the GostParc Sign corpus

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 2: GostParc signers: origin and gender

since there is no one-to-one correspondence between word and sign; also, Dutch and NGT syntax do not overlap. This corpus can serve as a publicly available training data evaluation set for the automatic translation of sign languages into spoken languages. As such we will set a benchmark in SLT.

The data will be soon made publicly available at the Instituut voor de Nederlandse Taal (INT) and through the European Language Grid under CC-BY licence. The metadata will also be in CMDI formats for harvesting by the CLARIN infrastructure. The EAMT logo will be prominently displayed on all Gost-Parc-Sign's material.

3 Gost-ParC-Sign - the project

The project has been developed in three phases.

All phases have been overseen by the Vlaams GebarenTaalCentrum (VGTC) and KU Leuven, both members of SignON, in order to ensure data and trans-

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 3: Age groups (at the time of recording)

lation quality.

1. Phase I was initiated by collecting the identified video material to be included in the corpus. In order to be able to share the corpus under CC-BY license,[7] we asked the owners of these videos to sign an informed consent form, which had obtained ethical clearance from the Research and Ethics Committee of Tilburg University.

2. For Phase II, which focused on manual translation, we hired a mixed team of deaf and hearing professional translators performing the task. Four translators worked in pairs consisting of one deaf and one hearing translator working together. This ensured that the original message of the VGT videos is preserved and that the Dutch text has good quality.

   Translations were created in ELAN (Sloetjes and Wittenburg, 2008). ELAN Annotation Format (EAF) files support multiple annotation tiers synchronised with the audio/video timeline. The "Translation" tier in each EAF file, corresponding to each video, is aligned at the sentence (or message) level with the VGT being signed.

   An example of data from the corpus is given in figure 1.

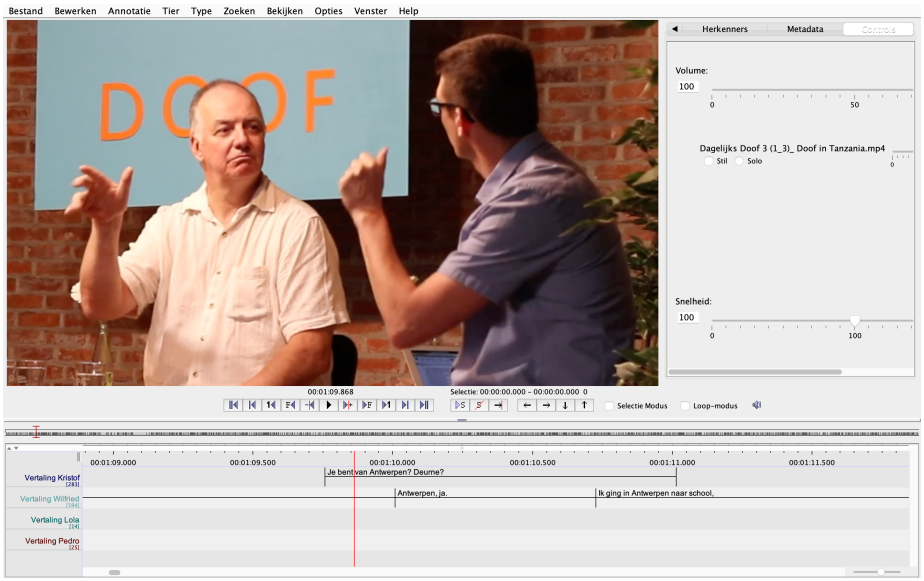

Figure 1: Example of ELAN annotation

The data in the two languages is aligned based on time stamps.

[7] in occasional circumstances, under CC-BY NC license

3. Once part of the translations were completed, phase III, the quality check phase started. This quality control was performed by members of the VGT deaf community and the VGTC and KU Leuven and by L1 Dutch users.

4 Conclusions

We have created a freely available corpus that can be used as a Gold Standard Parallel Corpus for translation between VGT and Dutch, with VGT as the source, as signed by L1 signers.

We hope that through creation of such a corpus the field of automated Sign Language translation has a clear benchmark dataset at its disposal onto which different approaches can be compared.

References

Camgoz, N. C., Hadfield, S., Koller, O., Ney, H., and Bowden, R. (2018). Neural sign language translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, USA. IEEE.

De Sisto, M., Shterionov, D., Murtagh, I., Vermeerbergen, M., and Leeson, L. (2021). Defining meaningful units. challenges in sign segmentation and segment-meaning mapping (short paper). In *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*, pages 98–103, Virtual. Association for Machine Translation in the Americas.

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. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2478–2487, Marseille, France. European Language Resources Association.

Graham, Y., Haddow, B., and Koehn, P. (2019). Translationese in machine translation evaluation. *CoRR*, abs/1906.09833.

Graham, Y., Haddow, B., and Koehn, P. (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, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Sloetjes, H. and Wittenburg, P. (2008). Annotation by category: ELAN and ISO DCR. In *Proceedings of the Sixth International Conference on Language*

Resources and Evaluation (LREC'08), Marrakech, Morocco. European Language Resources Association (ELRA).

Tiedemann, J. and Nygaard, L. (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. European Language Resources Association (ELRA).

Van Herreweghe, M., Vermeerbergen, M., Demey, E., De Durpel, H., Nyffels, H., and Verstraete, S. (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/.