A New English-Dutch-NGT Corpus for the Hospitality Domain

Mirella De Sisto
Tilburg University
Netherlands
M.DeSisto@tilburguniversity.edu

Vincent Vandeghinste
Instituut voor de Nederlandse Taal,
Leiden
Netherlands
KU Leuven, Belgium
Vincent.Vandeghinste@ivdnt.org

Dimitar Shterionov
Tilburg University
Netherlands
D.Shterionov@tilburguniversity.edu

Abstract

One of the major challenges hampering the development of language technology which targets sign languages is the extremely limited availability of good quality data geared towards machine learning and deep learning approaches. In this paper we introduce the NGT-Dutch Hotel Review Corpus (NGT-HoReCo), which addresses this issue by providing multimodal parallel data in English, Dutch and Sign Language of the Netherlands (NGT). The corpus contains 297 hotel reviews in written English (21.464 words), translated into written Dutch (22.274 words) and into NGT videos (230,54 minutes). It is publicly available through the ELG and the CLARIN platforms.

1 Introduction

As stated in Rivera Pastor et al. (2017), "The emergence of new technological approaches such as deep-learning neural networks, based on increased computational power and access to sizeable amounts of data, are making Human Language Technologies (HLT) a real solution to overcoming language barriers." Nevertheless, these very promising advances mainly concern HLT which focuses on spoken languages only, while HLT which targets sign languages is severely limited and strongly lagging behind (Vandeghinste et al., 2023).

This discrepancy between what has been achieved for spoken languages and what is available for signed languages is due to a number of challenges which are limiting the development of LT for signed languages (e.g. the lack of standardised data format, the lack of a standardised writing and annotation systems, etc.). For more details, see De Sisto et al. (2022).

The biggest bottleneck limiting the performance of new technological approaches for sign languages is the *quantity* of high quality data. To give an example, on average the data available for a relatively well resourced sign language is roughly ten times smaller than data available for a so-called low resource spoken language (Vandeghinste et al., forthcoming).

Besides the quantitative bottleneck, there is also an issue with data *quality*. Besides data scarcity, most of the parallel datasets which are available consist of spoken language news broadcasts interpreted into a sign language (in most cases by a hearing interpreter) (Camgoz et al., 2018). This affects the authenticity and the quality of the sign language data, since the interpreting process interferes with its accuracy (interpretation takes place simultaneously, which means that the interpreter needs to be quick and sometimes has to sacrifice accuracy for efficiency), and most hearing interpreters are not L1 users of the sign language (an exception being interpreters who are CODA — Children of Deaf Adults —and other specific cases).

The goal of the compilation of the NGT-Dutch Hotel Review Corpus (NGT-HoReCo) described in this paper is to contribute to reducing the scarcity of good quality sign language data by providing a multimodal parallel corpus of written English reviews and their translations into written Dutch and into NGT videos. The quality is ensured with respect to the authenticity of the NGT by the fact that translations were performed by deaf pro-

fessional translators. The accuracy of the translations is ensured by the fact that it concerns actual translations, performed in an offline modus without the constraints which are custom in an interpreting context.

The availability of a corpus such as NGT-HoReCo targets the stimulation of advancements in the field of sign language technology through both high-quality data for training models as well as a gold standard data for evaluation.

2 Related work

EASIER's Deliverable 6.1 (Kopf et al., 2021) and Morgan et al. (2022) provide an overview of the resources available for European sign languages.

NGT, together with German Sign Language (DGS), represent the richest sign languages in Europe in terms of available resources. Nevertheless, data available even for relatively well-represented sign languages are far from being sufficient for the development of language technologies.

The main source of data for NGT is the Corpus NGT (Crasborn et al., 2020), which is available for download at the Language Archive (`https://archive.mpi.nl/tla/`), in the form of separate files, and as a single file through the CLARIN infrastructure (`http://hdl.handle.net/10032/tm-a2-u5`). It contains 72 hours of dialogues between native users of NGT. 104 signers took part to the recordings. One limitation of the corpus is that only 25% of the data have been annotated (Crasborn et al., 2020); this is due to the fact that to date annotation is a manual and very-time consuming task (Morgan et al., 2022). As a consequence, only part of the Corpus NGT can be employed for MT tasks.

A different type of resource is constituted by lexicons. The lexicon of the Corpus NGT (Crasborn et al, 2020a) was made available by Global Signbank and is downloadable per sign. It consists of 3.645 short video files. Another available NGT lexicon downloadable per sign is `https://www.lerengebaren.nl/`, which consists of 2.993 videos.

3 Methodology: Preparation of the corpus

The creation of NGT-HoReCo required preparation of data for both Dutch and NGT. After gathering the publicly available English texts, these were translated into written Dutch; subsequently, the Dutch texts were translated offline into NGT videos by professional deaf translators.

3.1 Translation from English into Dutch

Written English is the source language of the hotel reviews from a Booking.com review corpus publicly available on Kaggle.[1] Reviews were selected with an initial manual screening which ensured that the texts were grammatically complete and correct, and that the text did not contain uncommon abbreviations. In some reviews with incomplete endings, final incomplete sentences were removed and the review was kept, when removal did not affect the meaning of the whole text; alternatively, the whole review ending in an incomplete sentence was removed.

The Dutch text side of the parallel corpus was produced by a professional translation company which used automatic translation (generated by DeepL) following and in-depth human post-editing.

The DeepL translations of the 297 reviews consists of 21.614 words, the post-edited version consists of 22.284 words.[2] An example entry is shown in Table 1.

3.2 Translation from Dutch into Sign Language of the Netherlands

The Dutch-NGT translation was performed by six professional deaf translators. The choice of having only deaf translators performing the task was made in order to ensure that the signing would be authentic and to reduce as much as possible the influence of the source language. For more details about why to use deaf translators, see Vandeghinste et al. (forthcoming). Translators were asked to sign an informed consent form which allows the data to be available under a CC BY-NC license.[3]

Each translation was recorded in a separate video file. Each review was translated once by a single translator.

An excel spreadsheet contains the written side of the parallel corpus: a column containing the English source, a column containing the DeepL translation, a column containing the post-edited version

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

[2] Calculated using the linux `wc` command.

[3] The project received ethical clearance from the Research Ethics and Data Management Committee of Tilburg University

Source	All in all the stay was good , but they were having issues with the elevator which was not good for being put on the 3rd floor
DeepL	Al met al was het verblijf goed, maar ze hadden problemen met de lift die niet goed was voor de 3e verdieping.
Video file	`NGT-HoReCo_1`
Post-edit	Al met al was het verblijf goed, maar ze hadden problemen met de lift, wat niet fijn is als je op de 3e verdieping wordt geplaatst.

Table 1: Example entry with its translation by DeepL and the post-edited version

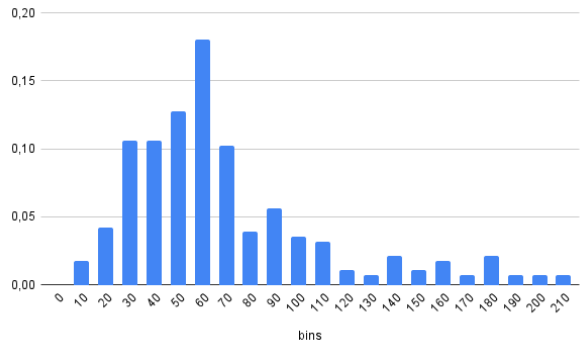

Figure 1: Length distribution of the post-edited Dutch translations, in bins of 10 words

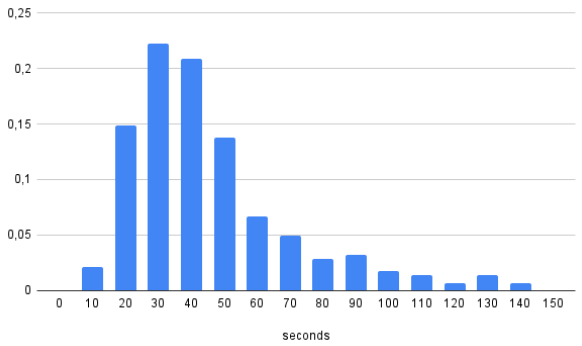

Figure 2: Length distribution of the video files in bins of 10 seconds

and a column containing the name of the corresponding NGT video file.

4 Results: NGT-HoReCo

The corpus comprises 297 hotel reviews (roughly 1.680 sentences) in written English, their translation into written Dutch and into NGT videos. The limited domain of the data, namely, 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.

The total amounts of words contained in the corpus is 21.464 for English and 22.274 for the Dutch text. The word length of the written reviews varies from around 15 to 400 words. The distribution of lengths in the post-edited translation is presented in Figure 1, where the X-axis is the length of the post-edited text, in bins of 10 words. The Y-axis is the ratio of files with a certain length.

The NGT translations consist of almost 4 hours of videos (230,54 minutes). The duration of the NGT videos ranges from around 10 seconds to around 4 minutes. The distribution of lengths of the videos is presented in Figure 2, where the Y-axis is the ratio of files and the X-axis is the duration in seconds, in bins of 10 seconds.

The corpus is publicly available through the ELG platform at `https://live.european-language-grid.eu/catalogue/corpus/21566`, and is also made available through the CLARIN platform. The permanent identifier for corpus download is `http://hdl.handle.net/10032/tm-a2-w2`. NGT-HoReCo is available under a CC BY-NC license, however, the written English text does not have availability restrictions, being fully publicly available in a Kaggle dataset.

5 Conclusion and future steps

In this paper we introduced a new available multimodal parallel corpus of written English, written Dutch and NGT videos. The corpus contains 297 hotel reviews in written English which were translated into written Dutch and into NGT videos. The Dutch-NGT translations were performed by deaf professional translators.

Parallel data such as NGT-HoReCo support further developments of Sign Language Technology, including but not limited to Sign Language Machine Translation.

A current limitation of the corpus is that there is no alignment between written sentences and video fragments. To date, there are no tools to automatically generate such alignment; consequently, a fur-

ther implementation of the corpus would include manual alignment.

In addition, the size of the corpus is still quite limited, due to time and cost restrictions of the NGT-HoReCo project.

Nevertheless, the advantage of the availability of parallel data such as NGT-HoReCo is that similar parallel corpora have the potential to be implemented with additional features and languages.

For instance, having the same reviews translated by more NGT translators coming from different parts of the Netherlands would account for language variation. We have considered this option but decided to first focus on having as many reviews translated as possible. Nevertheless, this would be a valuable direction for an implementation of the corpus.

Currently we have initiated a further development of NGT-HoReCo to also include Flemish Sign Language (VGT). Adding VGT is of particular interest because NGT and VGT, despite not being closely related languages, both base their mouthing on Dutch and are generally used in countries where Dutch is (one of) the official language(s). Additionally, NGT-HoReCo is going to be enriched with different types of annotations, such as pose estimates, etc.

Acknowledgments

The NGT-HoReCo project was funded by the European Language Equality 2 project (ELE 2), which has received funding from the European Union under the grant agreement no. LC-01884166 – 101075356 (ELE 2).

Work in this paper is part of the SignON project.[4] This project has received funding from 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.

Crasborn, O., I. Zwitserlood, J. Ros, and M. van Zuilen. 2020. Corpus ngt, 4e editie.

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.

Kopf, Maria, Marc Schulder, and Thomas Hanke. 2021. Overview of datasets for the sign languages of europe. Deliverable 6.1, Easier project.

Morgan, Hope E., Onno Crasborn, Maria Kopf, Marc Schulder, and Thomas Hanke. 2022. Facilitating the spread of new sign language technologies across Europe. In Efthimiou, Eleni, Stavroula-Evita Fotinea, Thomas Hanke, Julie A. Hochgesang, Jette Kristoffersen, Johanna Mesch, and Marc Schulder, editors, *Proceedings of the LREC2022 10th workshop on the representation and processing of sign languages: Multilingual sign language resources*, pages 144–147, Marseille, France. European Language Resources Association (ELRA).

Rivera Pastor, Rafael, Carlota Tarín Quirós, Juan Pablo Villar García, Toni Badia Cardús, and Maite Melero Nogués. 2017. Language equality in the digital age – Towards a Human Language Project. STOA study (PE 598.621), IP/G/STOA/FWC/2013-001/Lot4/C2, March 2017. Carried out by Iclaves SL (Spain) at the request of the Science and Technology Options Assessment (STOA) Panel, managed by the Scientific Foresight Unit (STOA), within the Directorate-General for Parliamentary Research Services (DG EPRS) of the European Parliament, March. `https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2017)598621`.

Vandeghinste, Vincent, Mirella De Sisto, Maria Kopf, Marc Schulder, Caro Brosens, Lien Soetemans, Rehana Omardeen, Frankie Picron, Davy Van Landuyt, Irene Murtagh, Elefterios Avramidis, and Mathieu De Coster. 2023. Report on Europe's Sign Languages. Technical report, European Language Equality D1.40.

Vandeghinste, Vincent, Mirella De Sisto, Santiago Egea Gomez, and Mathieu De Coster. forthcoming. Challenges with sign language datasets. In Way, Andy, Dimitar Shterionov, Lorraine Leeson, and Christian Rathmann, editors, *Sign Language Machine Translation*, chapter 10, pages 266–290. Springer, Oxford.

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