EUROPEAN LANGUAGE EQUALITY 2

FSTP Project Report

NGT-HoReCo – NGT-Dutch Hotel Review Corpus

Authors	Mirella De Sisto, Vincent Vandeghinste, Dimitar Shterionov
Organisation	Tilburg University, Instituut voor de Nederlandse Taal
Dissemination level	Public
Date	31-03-2023

About this document

Project	European Language Equality 2 (ELE2)
Grant agreement no.	LC-01884166 – 101075356 ELE2
Coordinator	Prof. Dr. Andy Way (DCU)
Co-coordinator	Prof. Dr. Georg Rehm (DFKI)
Start date, duration	01-07-2022, 12 months
FSTP Project	NGT-HoReCo – NGT-Dutch Hotel Review Corpus
Authors	Mirella De Sisto, Vincent Vandeghinste, Dimitar Shterionov
Organisation	Tilburg University, Instituut voor de Nederlandse Taal
Type	Report
Number of pages	8
Status and version	Final
Dissemination level	Public
Date of delivery	31-03-2023
EC project officer	Susan Fraser
Contact	European Language Equality 2 (ELE2) ADAPT Centre, Dublin City University Glasnevin, Dublin 9, Ireland Prof. Dr. Andy Way – andy.way@adaptcentre.ie European Language Equality 2 (ELE2) DFKI GmbH Alt-Moabit 91c, 10559 Berlin, Germany Prof. Dr. Georg Rehm – georg.rehm@dfki.de http://www.european-language-equality.eu © 2023 ELE2 Consortium

Consortium

1	Dublin City University (Coordinator)	DCU	IE
2	Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (Co-coordinator)	DFKI	DE
3	Univerzita Karlova (Charles University)	CUNI	CZ
4	Universidad Del Pais Vasco/ Euskal Herriko Unibertsitatea (University of the Basque Country)	UPV/EHU	ES
5	Athina-Erevnitiko Kentro Kainotomias Stis Technologies Tis Pliroforias, Ton Epikoinonion Kai Tis Gnosis	ILSP	GR
6	European Federation of National Institutes for Language	EFNIL	LU
7	Réseau européen pour l'égalité des langues (European Language Equality Network)	ELEN	FR

Contents

List of Figures

List of Tables

List of Acronyms

AI	Artificial Intelligence
AI4EU	AI4EU (EU project, 2019-2021)
CLAIRE	Confederation of Laboratories for AI Research in Europe
CLARIN	Common Language Resources and Technology Infrastructure
CRACKER	Cracking the Language Barrier (EU project, 2015–2017)
DARIAH	Digital Research Infrastructure for the Arts and Humanities
DLE	Digital Language Equality
EC	European Commission
ECSPM	European Civil Society Platform for Multilingualism
EFNIL	European Federation of National Institutes for Language
ELE	European Language Equality
ELE2	European Language Equality *(this project)*
ELE Programme	European Language Equality Programme *(the long-term, large-scale funding programme specified by the ELE project)*
ELEN	European Language Equality Network
ELEXIS	European Lexicographic Infrastructure
ELG	European Language Grid (EU project, 2019-2022)
ELRA	European Language Resource Association
ELRC	European Language Resource Coordination
ELT	European Language Technology
EP	European Parliament
ERIC	European Research Infrastructure Consortium
ESCO	European Skills, Competences, Qualifications and Occupations classification
GDPR	General Data Protection Regulation
KPI	Key Performance Indicator
LT	Language Technology/Technologies
META	Multilingual Europe Technology Alliance
META-NET	EU Network of Excellence to foster META
ML	Machine Learning
MT	Machine Translation
NCC	National Competence Centre
NCP	National Contact Point
NLP	Natural Language Processing
STOA	Science and Technology Options Assessment

Abstract

The major achievements of Language Technology (LT) nowadays mainly concern technology targeting spoken languages only. What is available for signed languages is instead extremely limited and strongly lagging behind. This is partly due to the fact that the amount of data available which can be used to support further developments in the field is scarce. With this project we aim to address this limitation by creating a multimodal parallel corpus of Dutch and Sign Language of the Netherlands (NGT). The data consists in hotel reviews in written Dutch on one side and their translation into NGT videos on the other side. The corpus will be made publicly available through the CLARIN and the ELG platforms. The goal of this corpus is to boost research into LT targeting sign languages and to support advances towards more inclusive language technology.

1 Introduction

As stated in STOA (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.” Such approaches are data-driven and require large amounts of linguistic data to train the parameters of the networks and reach good quality HLT.

Said language barriers do not solely exist between speakers of the 24 official European languages or speakers of other *spoken* languages. For approximately half a million of deaf and hard of hearing (DHH) people, sign languages are the main or preferred means of communication (Pasikowska-Schnass, 2018). Breaking down the language barriers between speakers and signers is therefore part of the European goal for language equality through technology by 2030.

HLT which targets sign languages is extremely limited. When we compare what is available and what has been achieved for spoken languages with the state of the art of technology for signed languages, we see a huge gap: the latter is severely lagging behind (Vandeghinste et al., 2023). Too often, when discussing HLT, sign languages are not even in the picture.

There are a number of challenges currently slowing down progress in HLT for sign languages (e.g. the lack of a widely accepted writing system for sign languages, the lack of standardised data format, etc.) (for more details, see De Sisto et al. (2022); Vandeghinste et al. (2023)). One of the major issues is the lack of good quality sign language data. Even when compared to the data available for low resource spoken languages, the data which is available on average for sign languages is much more limited (see Vandeghinste et al. (forthcoming)). For instance, the data available from one of the largest sign language corpora, the German Sign Language corpus (Prillwitz et al., 2008), is ten times smaller than the data available in Europarl (Koehn, 2005) for a low resource language (Vandeghinste et al., forthcoming).

In addition to that, most of the available parallel datasets consist in broadcasts in which the sign language is the result of interpretation (Camgoz et al., 2018). This affects the quality of the signing: firstly, interpretation is performed under time pressure, hence, the interpreter often needs to sacrifice the faithfulness to the original message in favor of efficiency; secondly, most hearing interpreters (made exception for CODA —Children of Deaf Adults —and other specific cases) do not use signing as their prominent means of communication, therefore, their fluency cannot be compared to that of an L1 user.

The NGT-Dutch Hotel Review Corpus (NGT-HoReCo) aims to reduce the good data availability gap by providing parallel data which can serve to support further development in LT targeting sign languages, be it as training data or as test data. NGT-HoReCo is a multimodal limited domain parallel corpus of Dutch text and Sign Language of the Netherlands (NGT) videos, in which NGT is the result of a translation performed by deaf professional translators.

2 Preparation of the corpus

NGT-HoReCo contains hotel reviews in written Dutch and their translation into NGT videos. The creation of the parallel corpus required gathering and preparation of data for both the Dutch and the NGT side. In the following sections, the activities carried out for both languages involved are described.

2.1 Dutch side

The Dutch text side of the parallel corpus was created by gathering hotel reviews, from a Booking.com review corpus publicly available on Kaggle, and by translating them into Dutch with DeepL. From this translated material, 350 reviews were selected based on the following criteria:

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

Some of the reviews contained final incomplete sentences. In those cases, we either removed the final incomplete sentences and kept the review, if removal did not affect the meaning of the whole text; alternatively, if the meaning would be modified by removing the final sentence, we excluded the review from the selection.

The selected 350 reviews were sent to a professional post editing company to ensure the good quality of the Dutch text.

Post editing took place in parallel with the translation tasks due to time limitations. However, we verified that the quality of the original DeepL translation did not display any severe problem.

2.2 NGT side

For the translation of the Dutch reviews into NGT, we contacted deaf professional translators, in order to reduce as much as possible the influence of the source language and to make sure that the signing is authentic. We hired three freelance translators and one translation company.

Since the data produced would consist of videos of the translators —hence, would inevitably contain personal information —before the beginning of the project, we had applied and obtained ethical clearance from the Research Ethics and Data Management Committee of Tilburg University. Translators were asked to sign an informed consent form with which they agreed with making this corpus publicly available.

We informed translators that we were aiming for good quality videos, made in normal, every-day life settings, hence preferably without employing blue/green screens.[1] Each video would need to contain one review and each review would be translated by only one translator. We considered the possibility of accounting for language variation by having the same

[1] Nevertheless, some of the videos have been provided with blue/green screen; given time restrictions, it was not possible to ask the translators to remake those videos.

reviews translated by different translators; however, given time and budget constraints, we decided to focus on having as many as possible reviews translated. Nevertheless, a possible follow-up project could focus on language variation.

In order to connect each video with its Dutch source text, we created an excel spreadsheet which contained each review and the name of the corresponding NGT video file.

3 NGT-HoReCo

The corpus comprises 283[2] reviews in written Dutch and their translation into NGT videos. The word length of Dutch reviews varies from around 15 to 400 words; the NGT videos duration ranged from around 10 seconds to around 4 minutes. The total amounts of words contained in the corpus is 21.825; the NGT translations consist of almost 4 hours of videos (213,18 minutes). The reviews have been translated by 6 deaf professional translators. The corpus is going to be made publicly available through the ELG and the CLARIN platform as soon as possible. At the moment it is available at https://b2drop.eudat.eu/s/HenFEKwAKMtzScT.

4 Summary and Conclusions

During this three month project, we created a Dutch-NGT parallel corpus. For this purpose 283 hotel reviews in written Dutch were translated into NGT videos by deaf professional translators. The availability of a similar corpus supports research focusing on more inclusive language technology, and in particular contributes to the efforts towards language technology which also targets sign languages.

References

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, USA, 18 – 22 June 2018. IEEE. URL http://personal.ee.surrey.ac.uk/Personal/S.Hadfield/papers/camgoz18cvpr.pdf.

Mirella De Sisto, Vincent Vandeghinste, Santiago Egea Gómez, Mathieu De Coster, Dimitar Shterionov, and Horacio Saggion. 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 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.264.

Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers, MTSummit 2005, Phuket, Thailand, September 13-15, 2005*, pages 79–86, 2005. URL https://aclanthology.org/2005.mtsummit-papers.11.

Magdalena Pasikowska-Schnass. Sign languages in the EU. Technical report, European Parliamentary Research Service, 2018. URL http://www.europarl.europa.eu/RegData/etudes/ATAG/2018/625196/EPRS_ATA(2018)625196_EN.pdf.

Siegmund Prillwitz, Thomas Hanke, Susanne König, Reiner Konrad, Gabriele Langer, and Arvid Schwarz. DGS corpus project–development of a corpus based electronic dictionary German Sign Language/German. In *3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora*, page 159, 2008.

[2] The original amount of reviews selected was 350; however, given the time limitation, this is the amount of translations that the translators managed to perform.

STOA. 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 2017. http://www.europarl.europa.eu/stoa/.

Vincent Vandeghinste, 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. Report on Europe's Sign Languages. Technical report, European Language Equality D1.40, 2023.

Vincent Vandeghinste, Mirella De Sisto, Santiago Egea Gómez, and Mathieu De Coster. Challenges with sign language datasets, forthcoming.