SignON – a Co-creative Machine Translation for Sign and Spoken Languages (end-of-project results, contributions and lessons learned)

Dimitar Shterionov[*], **Vincent Vandeghinste**[†a], **Mirella De Sisto**[*], **Aoife Brady**[‡], **Mathieu De Coster**[§], **Lorraine Leeson**[¶], **Josep Blat**[**], **Frankie Picron**[††], **Davy Van Landuyt**[††], **Marcello Paolo Scipioni**[‡‡], **Andy Way**[‡], **Aditya Parikh**[§§], **Louis ten Bosch**[§§], **John O'Flaherty**[‖], **Joni Dambre**[§], **Caro Brosens**[x], **Jorn Rijckaert**[x], **Bram Vanroy**[a], **Victor Ubieto Nogales**[**], **Santiago Egea Gomez**[**], **Ineke Schuurman**[a], **Gorka Labaka**[b], **Adrián Núñez-Marcos**[b], **Irene Murtagh**[c], **Euan McGill**[**], **Horacio Saggion**[**]

[*]Tilburg University, [†]Instituut voor de Nederlandse Taal, [‡]ADAPT, [§]Ghent University, [¶]Trinity College Dublin,[**]Universitat Pompeu Fabra,[††]European Union of the Deaf, [‡‡]Fincons,[§§]Radboud University,[‖]mac.ie,[x]Vlaams Gebarentaalcentrum, [a]KU Leuven, [b]University of the Basque Country UPV/EHU, [c]TU Dublin

SignON,[1] a 3-year Horizon 2020[2] project addressing the lack of technology and services for MT between sign languages (SLs) and spoken languages (SpLs) ended in December 2023. SignON was unprecedented. Not only it addressed the wider complexity of the aforementioned problem – from research and development of recognition, translation and synthesis, through development of easy-to-use mobile applications and a cloud-based framework to do the "heavy lifting" as well as to establishing ethical, privacy and inclusiveness policies and operation guidelines – but also engaged with the deaf and hard of hearing communities in an effective co-creation approach where these main stakeholders drove the development in the right direction and had the final say.

Currently we are witnessing advances in natural language processing for SLs, including MT. SignON was one of the largest projects that contributed to this surge with 17 partners and more than 60 consortium members, working in parallel with other international and European initiatives, such as project EASIER[3] and others.

SignON MT – framework SignON set out to develop an MT service supporting 4 SpLs (Dutch, Spanish, Irish and English) in both written and spoken forms and 5 SLs (Sign Language of the Netherlands (NGT), Flemish Sign Language (VGT), Spanish Sign Language (LSE), Irish Sign Language (ISL) and British Sign Language (BSL)) in any possible direction. A fleet of dedicated language-pair-specific models would be infeasible. Considering the unsustainable nature of such an approach SignON employed a divide-and-conquer strategy splitting the task into automatic speech recognition (ASR) and sign language recognition (SLR), MT and synthesis. The MT core we built (i.e. the *InterL*) is based on mBART (Lewis et al., 2020) and on symbolic representations and aims to capture the meaning of all languages (spoken and signed) and to facilitate the translation processes; ASR and SLR components provide input to the InterL – text in the case of ASR and visual and temporal embeddings for SLR; text outputs from InterL is displayed to the user or fed into an SLS and a text-to-speech component to generate the target language utterance in the targeted modality.

SignON – Co-creation Up till SignON commenced, SLMT work lacked the proper inclusion of deaf and hard of hearing people in the process of planning projects, participating as equal partners in researching, and responding to work in development stage (Bragg et al., 2019).

To address the aforementioned gap, the SignON project involved the deaf and hard of hearing communities from the beginning. First, two deaf-led organisations were involved from the inception stage with leading roles within the consortium. Second, SignON employed a co-creation approach which places deaf and hard of hearing stakeholders at the centre of the design process. We defined co-creation as *a collaboration between researchers, developers and users, based on continuous, periodic information exchange, expectation management, openness and user-involvement (in the design and development process), on equal merits and built on trust, from the project inception*. We conducted 12 co-creation events and surveys spread over duration of the project and ge-

[1]https://signon-project.eu/

[2]RIA Grant Agreement No. 101017255

[3]https://project-easier.eu/

ographically over the regions of every involved language. Co-creation events included interviews, round tables and workshops. Following analysis of the collected feedback, information was fed into the development cycle which, ultimately, led to 3 major releases (of the SignON MT App).

SignON – an open and sustainable solution The SignON services involve various models, tools and components that are deployed on a distributed framework. Code, models and documentation are available on github (`https://github.com/signon-project`) and huggingface (`https://huggingface.co/signon-project`). Public deliverables, which describe the SignON outputs and outcomes are available at `https://signon-project.eu/publications/public-deliverables/`. The availability of code, models, documentation and data is only one side of the sustainability dimension of SignON. Alongside the translation pipelines, we have embedded a learning/training aspect. In particular, (i) we built pipelines to easily adapt models to new data and (ii) more importantly, alongside the translation app we developed a data collection app which allows the collection of user-generated data.

The wide span of SignON led to many satellite initiatives. Two of these, ELE- and EAMT-funded projects (De Sisto et al., 2023a; De Sisto et al., 2023b), focused on data collection, as data (or the lack of it) is one of the main challenges uncovered in SignON. In particular, the data-related problems include insufficient (for deep learning) volumes of data, formatting and difficulties with processing of data,[4] quality of the data (Vandeghinste et al., forthcoming), and even related to data authenticity i.e. most of the data is generated by hearing interpreters who can be considered L2 signers but also translating under time pressure, leading to "translationese". These two projects led to the generation of two CC-BY NC and CC-BY licensed datasets, available or soon to be available through CLARIN[5] and the ELG,[6] aiming to further the development of NLP for SLs. Other SignON data is distributed through the CLARIN infrastructure and is listed in: `https://www.clarin.eu/resource-families/sign-language-resources`.

[4]As noted in (Morgan et al., 2022; Vandeghinste et al., 2023) while the majority of the SL data is stored as videos, no automatic annotation tool is available, requiring manual work.

[5]`http://hdl.handle.net/10032/tm-a2-x5`, `http://hdl.handle.net/10032/tm-a2-x4`, `http://hdl.handle.net/10032/tm-a2-x6`

[6]`https://live.european-language-grid.eu/catalogue/corpus/21535`, `https://live.european-languagegrid.eu/catalogue/corpus/23007`

References

Bragg, Danielle, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, Christian Vogler, and Meredith Ringel Morris. 2019. Sign language recognition, generation, and translation: An interdisciplinary perspective. In *ASSETS 2019 – 21st International ACM SIGACCESS Conference on Computers and Accessibility*, pages 16–31.

De Sisto, Mirella, Dimitar Shterionov, Lien Soetemans, Vincent Vandeghinste, and Caro Brosens. 2023a. Ngt-horeco and gost-parc-sign: Two new sign language - spoken language parallel corpora. In Krister Lindén, Jyrki Niemi and Thalassia Kontino, editors, *CLARIN Annual Conference Proceedings*, Leuven, Belgium.

De Sisto, Mirella, Vincent Vandeghinste, Lien Soetemans, Caro Brosens, and Dimitar Shterionov. 2023b. GoSt-ParC-sign: Gold standard parallel corpus of sign and spoken language. In Nurminen, Mary, Judith Brenner, Maarit Koponen, Sirkku Latomaa, Mikhail Mikhailov, Frederike Schierl, Tharindu Ranasinghe, Eva Vanmassenhove, Sergi Alvarez Vidal, Nora Aranberri, Mara Nunziatini, Carla Parra Escartín, Mikel Forcada, Maja Popovic, Carolina Scarton, and Helena Moniz, editors, *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 503–504, Tampere, Finland, June. European Association for Machine Translation.

Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In D. Jurafsky et al., editor, *Proc. of the 58th Annual Meeting of the Assoc. for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. ACL.

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).

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 Gómez, and Mathieu De Coster. forthcoming. Challenges with sign language datasets.