LREC 2022

Challenges with Sign Language Datasets for Sign Language Recognition and Translation

Mirella De Sisto[1], Vincent Vandeghinste[2], Santiago Egea Gómez[3], Mathieu De Coster[4], Dimitar Shterionov[1], Horacio Saggion[3]
[1] Tilburg University, [2] Instituut voor de Nederlandse Taal, [3] Universitat Pompeu Fabra, [4] Ghent University

Context:

NNs have propelled the research on automatic MT systems. However, most advancements concern its use for spoken languages; while MT applied to SLs is lagging far behind. This discrepancy is partially due to a number of challenges which are related to SL data.

Challenges:

1) Data is limited and sparse.
2) The source language of available data might not be authentic SL (i.e. data quality).
3) Acquiring the data as downloadable datasets is often difficult.
4) Annotations vary in terms of format, type, and granularity
5) The data formats have limited usability for MT.

	Publicly available	Format	Source Lang.	Signers per file	R & L hand glosses	Non-manual features	Available annotation guidelines	OpenPose
ECHO Corpus	separate files	ELAN	BSL	1	2	✓	✓	✗
BSL Corpus	separate files	ELAN	BSL	2	2	✗	✓	✗
Corpus VGT	separate files	ELAN	VGT	2	2	✓	✓	✗
Corpus NGT	separate files	ELAN	NGT	2	2	✓	✓	✗
iSignos	on website	CSV	LSE	1	2	✗	✓	✗
Porta's corpus	thesis	-	ES	1	1	✗	✓	✗
Signs of Ireland	private	ELAN	ISL	1	1	✓	✗	✗
Content4All	with account	JSON	NL	1	n/a	n/a	n/a	✓
RWTH-PHOENIX	yes	CSV	DE	1	1	✗	✗	✗
How2Sign	not fully	CSV	EN	1	1	-	✗	✓
CNSE's corpus	only video	-	ES	-	-	-	-	-
LSC corpus	✗	iLex	CAT	1	2	✓	✓	✗
BOBSL corpus	restricted	-	EN	-	n/a	✗	n/a	✗

Differences in format and annotated data

Corpus	Pointing to the signer
BSL Corpus [1]	PT:PRO1SG
Corpus NGT [2]	IK
Corpus VGT [3]	WG-1
iSignos [4]	INDX.PRO:1sg
Spanish-LSE corpus [5]	YO

Differences in glossing conventions

Abbrevations: Neural Network (NN), Sign Language (SL), Machine Learning (ML), Machine Translation (MT).

*The **SignON project** focuses on the research and development of a SL translation mobile application and an open communications framework.*

Current proposal:
A framework to unify ELAN data into MT-suitable format

(1) ELAN file parsing

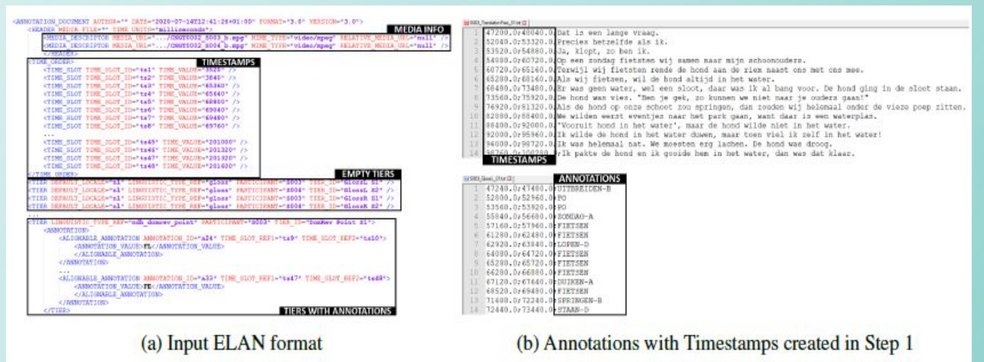
Data in the annotation tiers are extracted, together with the media information and annotations timestamps. Empty tiers are skipped, and different participants (if more than one) recognised (a). The annotations along with their timestamps are stored in the single text folders as CVS-like format (b), whereas media data are stored in files in the folders created for each ELAN file.

(2) Aligning and merging annotations

The user must define one leading modality and other required modalities to process. Then, the timestamps related to the leading modality will be used to align the other modalities, and create aligned sequences (c).

(3) Video frames extraction

When videos are matched with participants, the files generated in (1) are used to check the timestamps in which signs are being produced. Those video frames are extracted, resized (224x224) and stored in the subfolder *videoframes*.



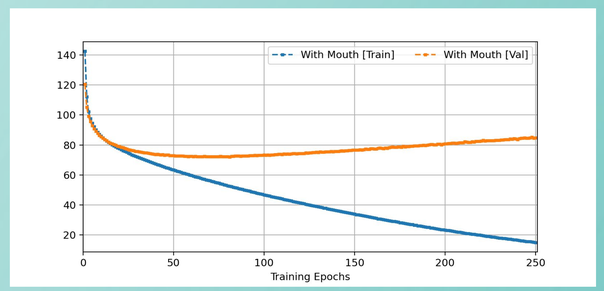
(a) Input ELAN format

(b) Annotations with Timestamps created in Step 1

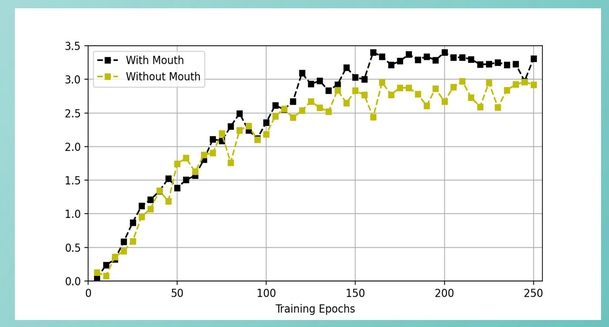
(c) Aligned Text created in Step 2

Experimental setting (Corpus NGT)

- Leading modality: *Free Translation* tier
- Required modalities: *GlossL, GlossR* and *Mouth*
- Input of the transformer: *GlossL, GlossR* and *Mouth*
- Output of the transformer: *Free translation*
- Two MT models: one with all inputs and one without mouthing



Train and validation losses for model with mouthing



BLEU scores on the validation partition for the models with and without mouthing.

[1] Schembri, A. and Fenlon, J. and Rentelis R. and Stamp, R. and Cormier, K. (2011). *British Sign Language Corpus Project.* [2] Crasborn, O. and Zwitserlood, I. and Ros, J. and van Zuilen, M. (2020). *Corpus NGT, 4e editie.* [3] Van Herreweghe, M. & Vermeerbergen, M. and Demey, E. and De Durpel, H. and Verstraete, S. (2015). *Het Corpus VGT. Een digitaal open access corpus van videos and annotaties van Vlaamse Gebarentaal, ontwikkeld aan de Universiteit Gent i.s.m. KU Leuven.* [4] Cabeza, C. & J. M. García-Miguel. (2018). *iSignos: Interfaz de datos de Lengua de Signos Española (version 1.0).* [5] Porta, J. (2014). *Towards a rule-based Spanish to Spanish sign language translation: from written forms to phonological representations.* Ph.D. thesis, Universidad Autónoma de Madrid. Departamento de Tecnología Electrónica y de las Comunicaciones.

The SignON project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017255