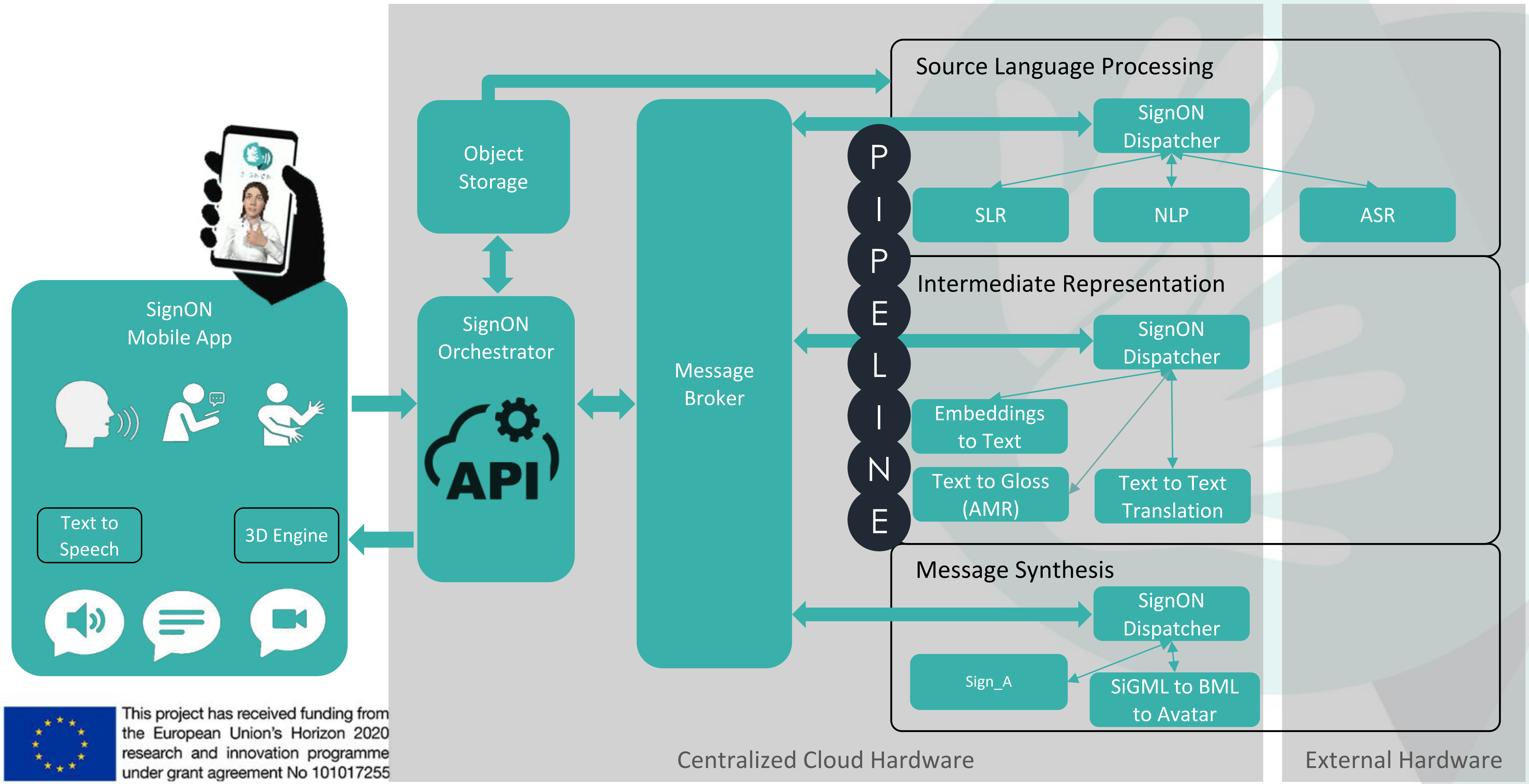
The SLMT pipeline

SignON and EASIER Final Event
29/ 11 /2023, Brussels

Cloud Platform Architecture



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

Input: a sign language video

SLR

extract information from input videos containing signed utterances and process it into a suitable format for any downstream task

MT

convert an input sequence in one language (the source) into another language (target) while preserving its meaning

SLS

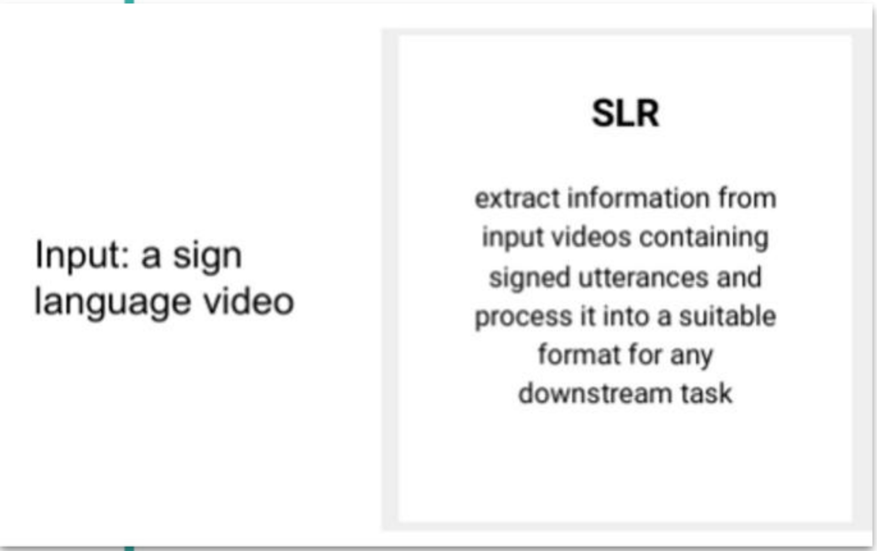
generate signed utterances in the target sign language.

Output: an animation or a video representing a sign language utterance

SLR output/MT input: a symbolic or vectorial representation of signs as tokens

MT output/SLS input: a spatio-temporal representation of signs that could be rendered (or interpreted) as a signed utterance.

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

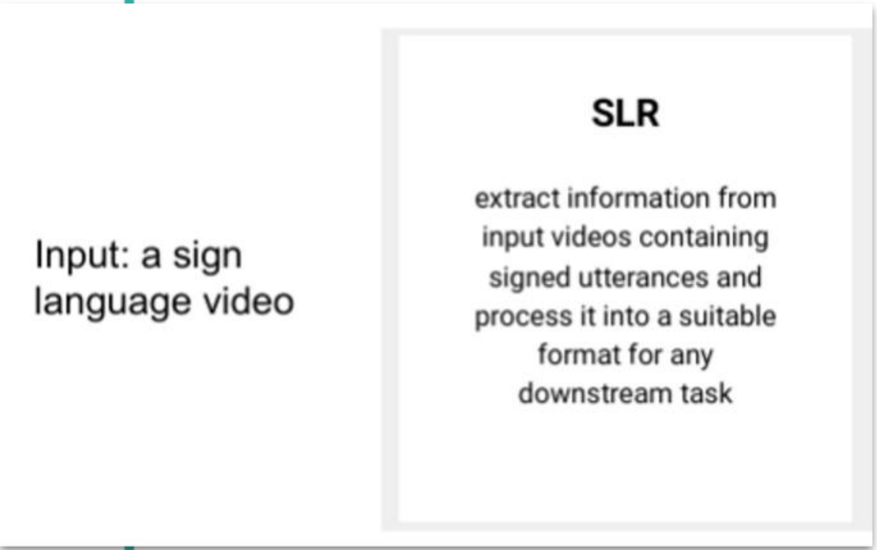


SLR is the task of extracting information from an input video containing signs, and processing this information into a form suitable for any downstream task, e.g. MT

End-to-end, as with most AI tasks is the preferred way. However, no sufficient high-quality annotated data =>

feature extractor → sign language classifier

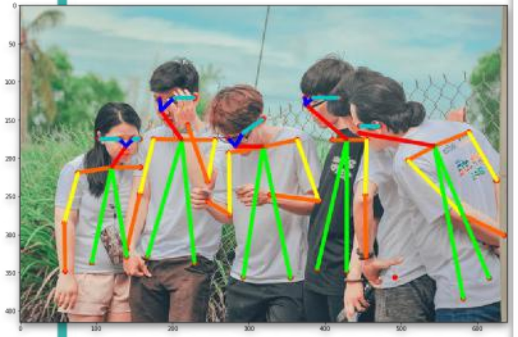
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017255



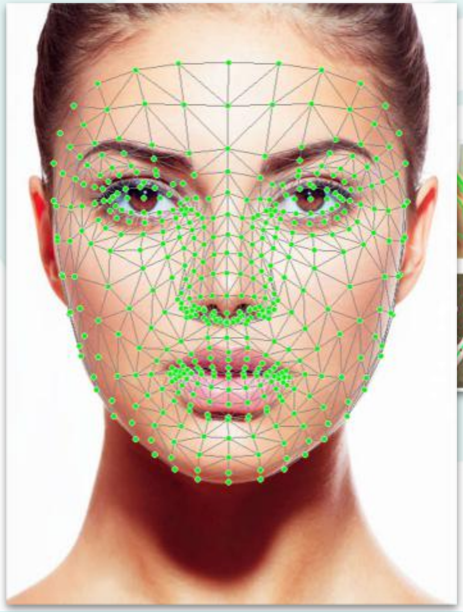
SLR is the task of extracting information from an input video containing signs, and processing this information into a form suitable for any downstream task, e.g. MT

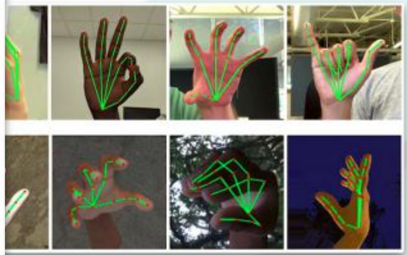
End-to-end, as with most AI tasks is the preferred way.
However, no sufficient high-quality annotated data =>

feature extractor → sign language classifier

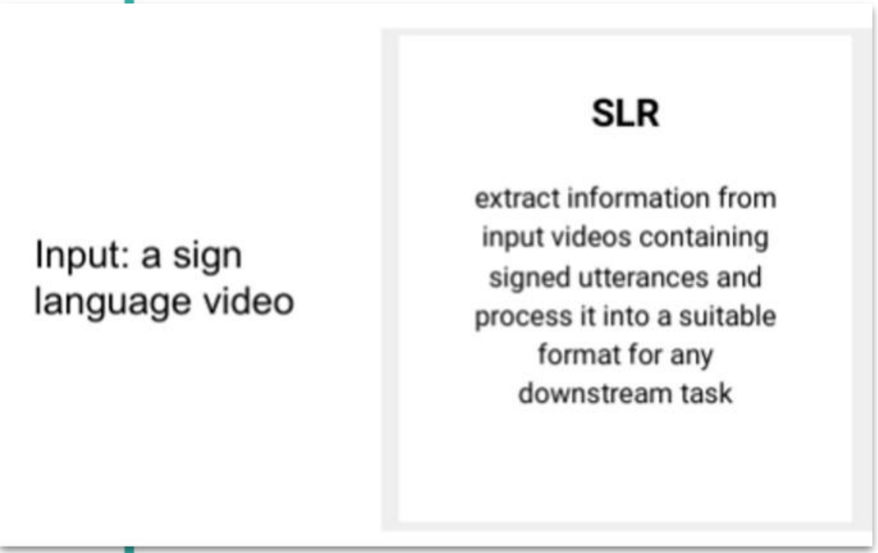


OpenPose





This project has receive[d] ... the European Union's ... research and innovation programme under grant agreement No 101017255



SLR is the task of extracting information from an input video containing signs, and processing this information into a form suitable for any downstream task, e.g. MT

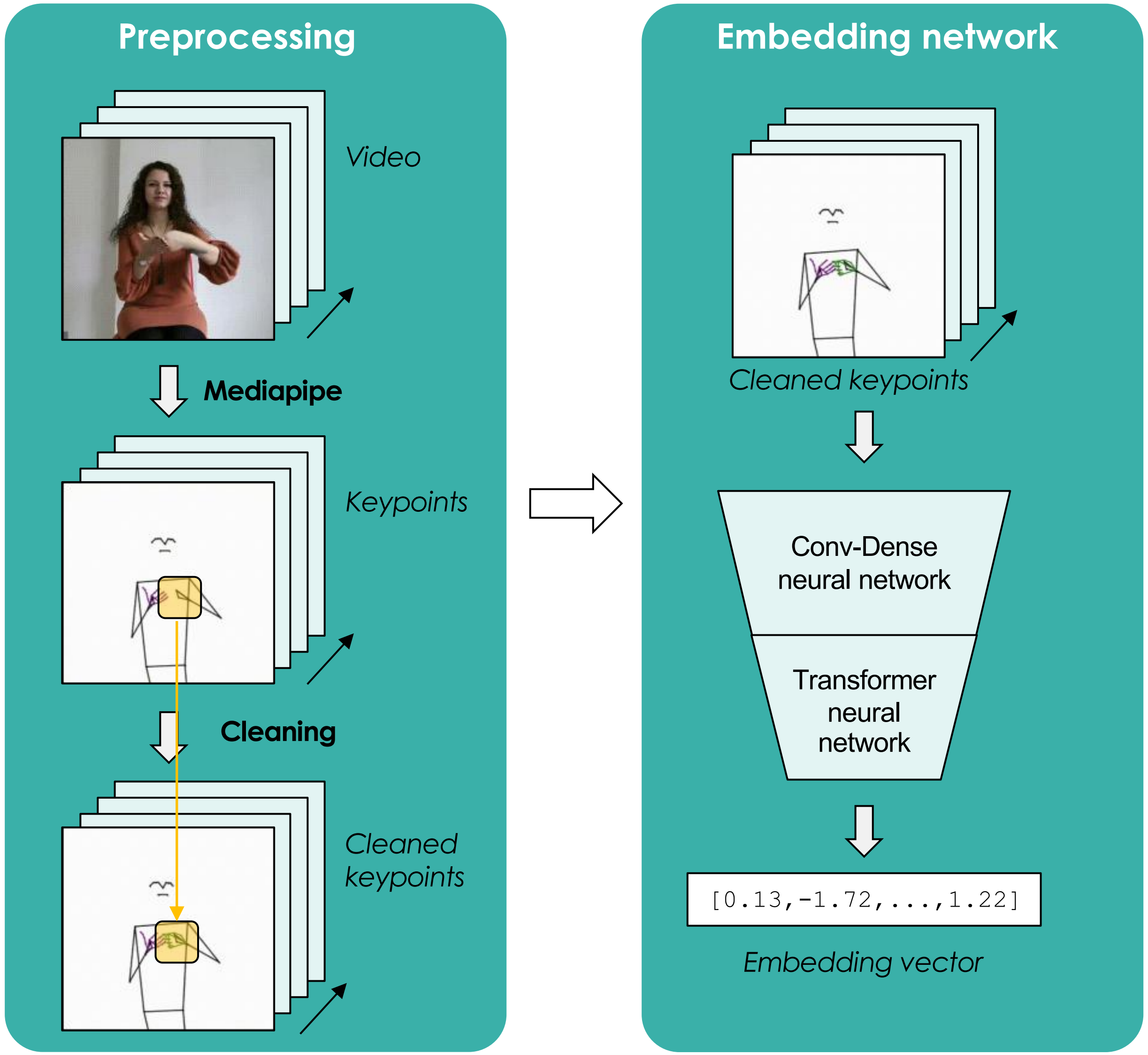
Isolated SLR = each data example corresponds to a single sign; the objective is to learn how to classify each such sample.

Continuous SLR = data samples contain one or more signs, i.e. a video of continuous signing. The task is then to both locate and recognize signs deriving a sequence of representative tokens.

Train an isolated SLR ⇒
extract SL representations from the data ⇒
use these representations for downstream tasks such as SLMT

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

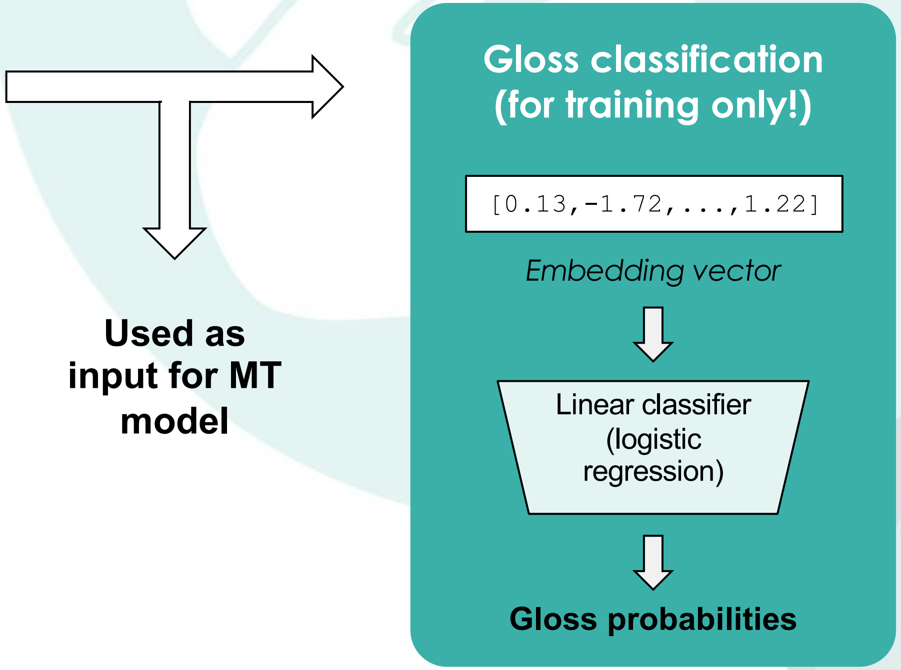
SLR model and training



Available gloss-labeled training data

	VGT	NGT	ISL	BSL
Samples (sign videos)	24967	68854	4103	2635
Classes (different glosses)	292*	458*	224**	124**
Different signers	111	82	37	48

** : only glosses with at least* ***20 examples***
: only glosses with at least* ***5 examples



SLR: evolution since mid-term review

New model was developed*:

> powerful/efficient model architecture
> significant improvement for all SLs

Mitigating limited data set size:

- multi-language pre-training + fine-tuning for each language
- pre-training on largest dataset (NGT),
  only training classification head for target language

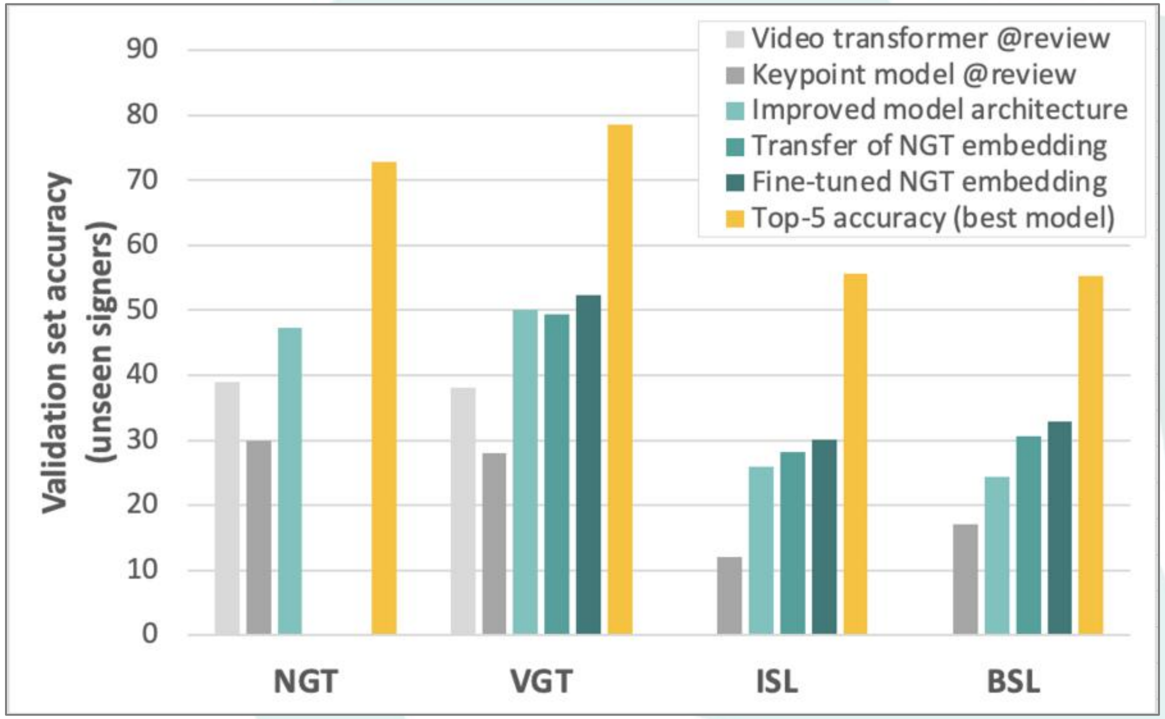
> improvement for ISL/BSL
> worse for VGT

- pre-training on largest dataset (NGT),
  transfer & fine tuning for target language

> Best results thus far!

*: Kaggle competition (ASL-ISLR):

> 94,000 samples with 250 glosses
> SignON ranked 16th/1165 - 86.6% accuracy



Top-5 accuracy

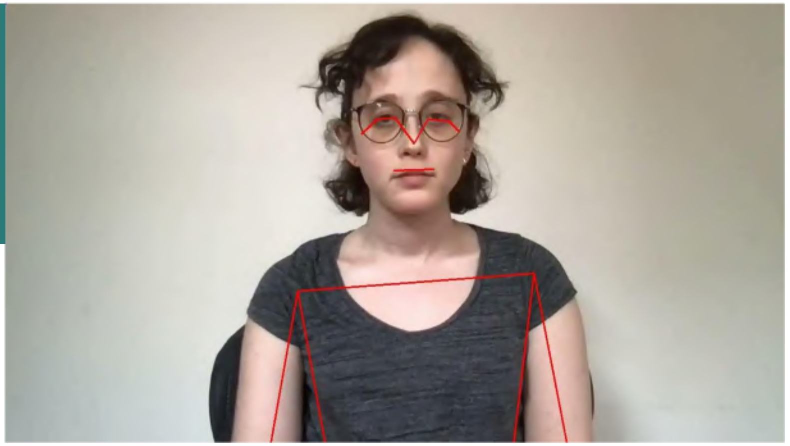
Very high for NGT/VGT
Embeddings contain relevant information about gloss
Mistakes are often confusions between similar glosses

SLR model predictions

Only glosses with predicted probability >0.4 kept

red = wrong, ***light green cursive*** = similar, **dark green** = correct

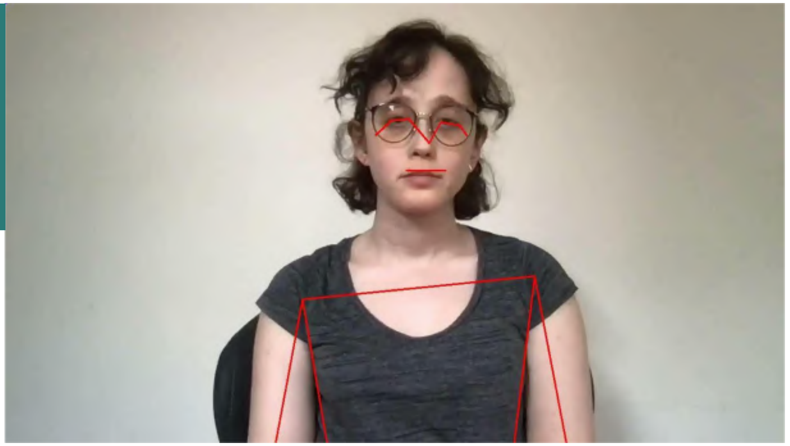
Example 1



Annotation: SCHILDPAD EERST AANKOMEN

Prediction: **SCHILDPAD EERST AANKOMEN**

Example 2



Annotation: WANNEER MOETEN AUTO HALEN WETEN-NIET IK

Prediction: **WANNEER WG-3 MAAR ROLLEN *AUTO-RIJDEN* *PAKKEN* WG-3 OK WETEN-NIET IK**

Example 3

Annotation: WAT WILLEN ETEN JIJ

Prediction: VINDEN **WAT *<missing>* *<missing>* JIJ** ZOEKEN MOOI

Testing on newly recorded phrases:

- **Example 1:** perfectly predicted
- **Example 2:**
  some wrong glosses instead of missed gloss *(MOETEN),*
  some glosses not exact but very close
- **Example 3:**
  some spurious glosses (start and end of recording)
  some glosses missed (each *predicted as 3rd option*)

-> analyses/user feedback will be used to make further improvements

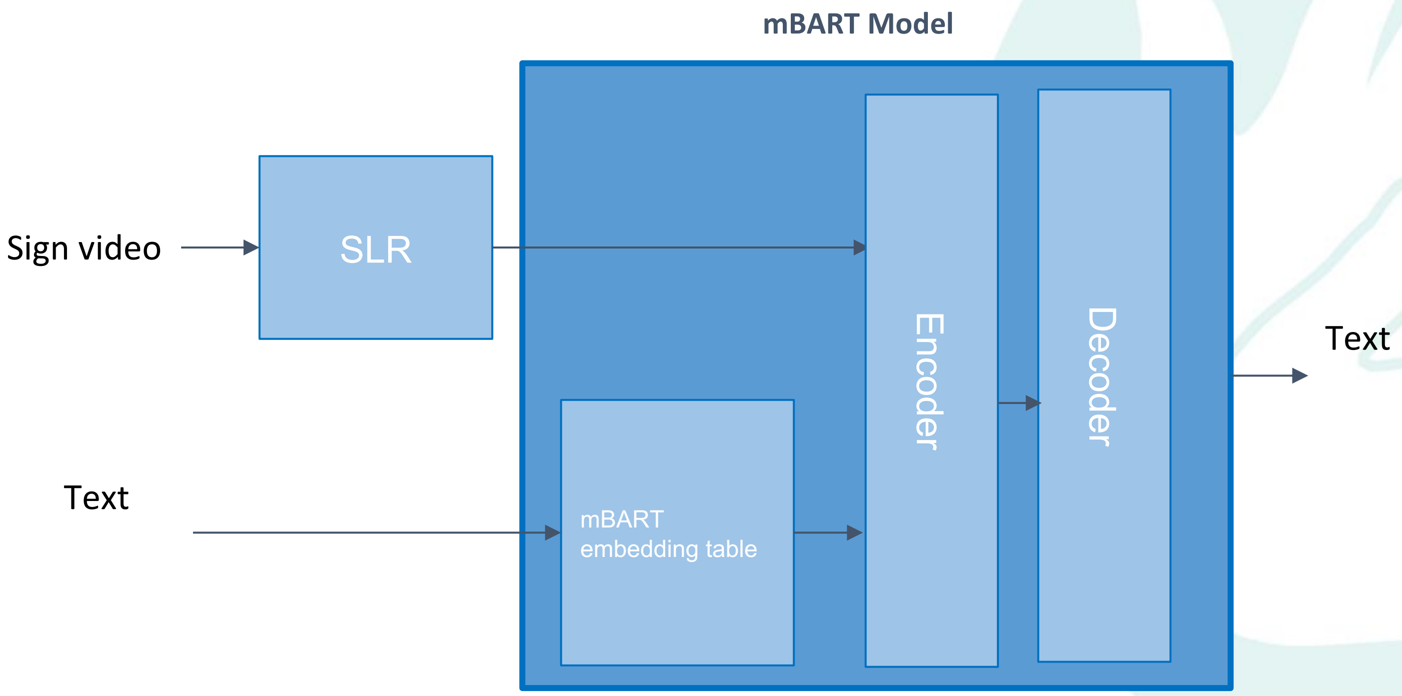
MT

convert an input sequence in one language (the source) into another language (target) while preserving its meaning

Machine translation is the process of automatically translating content from one language (the source) to another (the target) without human intervention.

RBTM	(PB)SMT	NMT - RNN	NMT - Transformer	NOW
Human-crafted rules. Difficult to update. But can work (relatively well) for low-resource languages.	Data-driven. Phrase-based translation probabilities. Translation and language models. Decoder. Noisy Channel. First commercial breakthroughs. Has reached its limits.	Encoder-decoder architecture. LSTMs and attention to solve problems with long sequences. One token after another. Cannot be parallelised.	Self-attention. Positional encoding. Feed-forward networks. Can be parallelised. Efficient and high quality. Mimics the way humans would translate a sentence.	BERT, ELMO, GPT, XLM, NLLB Unsupervised models. Better data preprocessing. Multi-lingual, multimodal models. Pay more attention to linguistics. Control bias.

Machine Translation Module



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

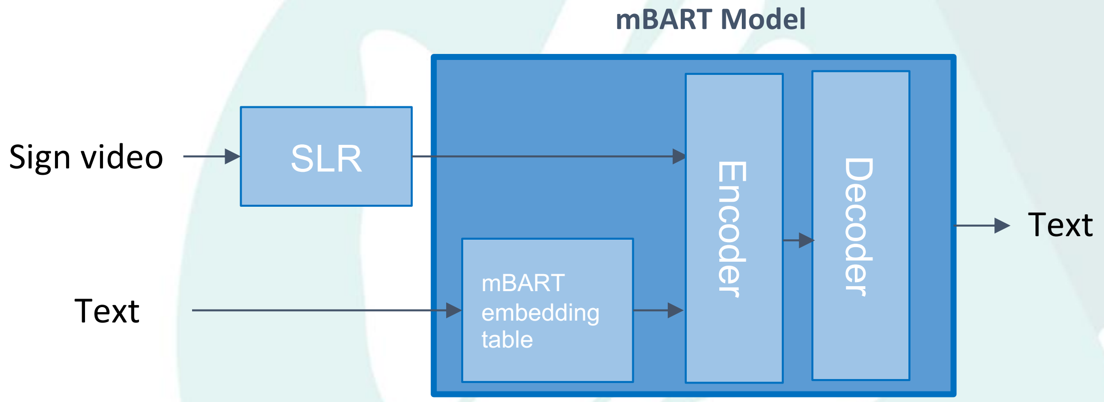
[1] Camgöz, N. C., Saunders, B., Rochette, G., Giovanelli, M., Inches, G., Nachtrab-Ribback, R., & Bowden, R. (2021, December). Content4all open research sign language translation datasets. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)* (pp. 1-5). IEEE.

Machine Translation Module

Training and evaluation data

Text-to-text data:
- Paracrawl 7.1 (SignON spoken languages)

SL-to-text Data:
- Content4All VRT-NEWS (VGT-Dutch) [1]
- Phoenix2014T (DGS-German)



Results

Text-to-text results on Paracrawl						
Metric / Data	EN-GA	GA-EN	EN-ES	ES-EN	EN-NL	NL-EN
BLEU	48.56	55.38	40.68	41.21	43.97	48.75
TER	0.50	0.46	0.50	0.51	0.59	0.55

Sign-to-text results		
Metric / Data	Phoenix2014T	Content4All VRT-NEWS
BLEU	22.66	0.44
CHRF	48.58	16.5
ROUGE	43.86	8.94

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

[1] Camgöz, N. C., Saunders, B., Rochette, G., Giovanelli, M., Inches, G., Nachtrab-Ribback, R., & Bowden, R. (2021, December). Content4all open research sign language translation datasets. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)* (pp. 1-5). IEEE.

Machine Translation Module

SLT results on Content4All's VRT-NEWS VGT-Dutch dataset

- Much lower than SLT results on Phoenix2014T.
- Similar to the SLT results obtained…
  - in the original paper.
  - in the WMT shared task (although the dataset is not the same) [2].

	BLEU		
Submission	all	SRF	
UZH (baseline)	0.12±0.06	0.09±0.03	0.
DFKI-SLT	0.08±0.01	0.10±0.04	0.
DFKI-MLT.1	0.07±0.05	0.05±0.02	0.
DFKI-MLT.2	0.11±0.06	0.08±0.03	0.
DFKI-MLT.3	0.08±0.04	0.06±0.02	0.
DFKI-MLT.4	0.02±0.01	0.02±0.01	0.
DFKI-MLT.5	0.04±0.02	0.03±0.00	0.
MSMUNICH.1	0.44±0.21	0.34±0.18	0.
MSMUNICH.2	0.56±0.30	0.28±0.13	0.
NJUPT-MTT.1	0.09±0.01	0.13±0.03	0.
NJUPT-MTT.2	0.10±0.01	0.13±0.03	0.

Reasons for the low results

- Insufficient amount of data (~7,000 video-sentence pairs for training).
- Phoenix2014T has a narrower domain (weather forecast), smaller vocabulary (~2,000-3,000 subwords).
  - Moreover, glosses were used in the training.
- The domain of the datasets used for SLR and SLT do not match.

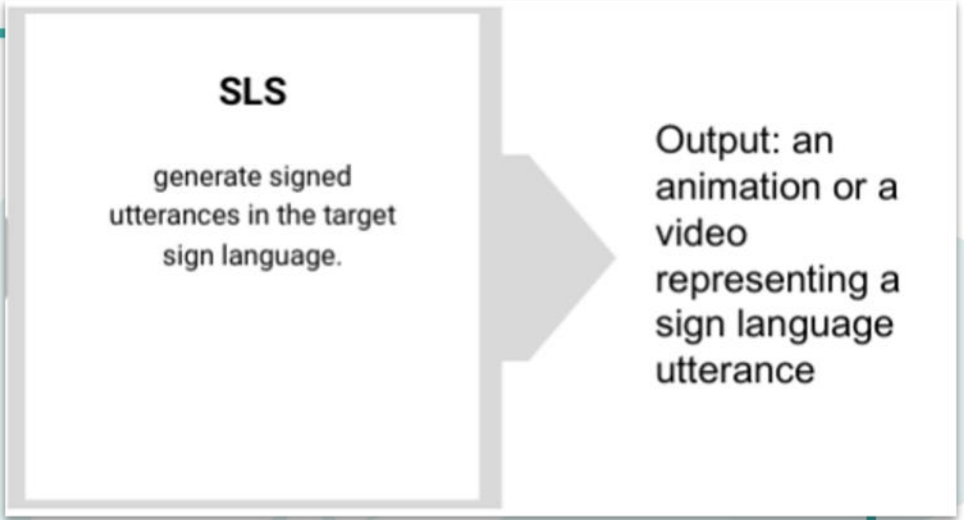
Conclusions

- Results comparable to the state of the art.
- Continuous model / architectural improvements

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

[2] Mathias Müller et al. (2022) [Swiss German Sign Language -> German]
Findings of the First WMT Shared Task on Sign Language Translation (WMT-SLT22)

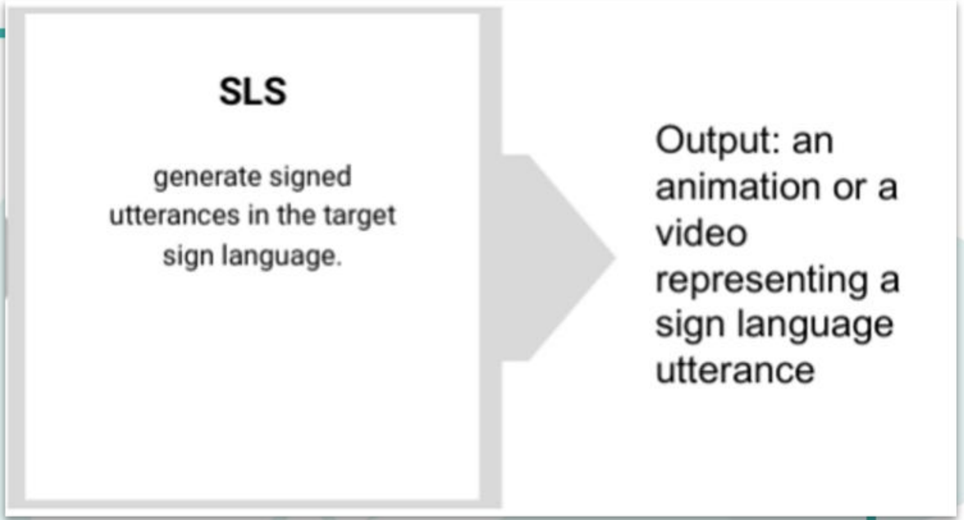
Sign language synthesis (SLS) or sign language production is the task of generating a synthetic representation that can exhibit properties of a human signer and utter a message in a SL through the expression of manual features (hand configuration, location, and orientation) and non-manual features (including facial expressions, mouthing and mouth gestures, gaze and torso direction).



- 3D animation-based approach which resolves in generating a 3D animated character, commonly referred to as an avatar
- a (video of a) virtual human that can be synthesised with generative AI methods based on real human video/image data (see the work of Stoll et al. (2020)).

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

Sign language synthesis (SLS) or sign language production is the task of generating a synthetic representation that can exhibit properties of a human signer and utter a message in a SL through the expression of manual features (hand configuration, location, and orientation) and non-manual features (including facial expressions, mouthing and mouth gestures, gaze and torso direction).



- 3D animation-based approach which resolves in generating a 3D animated character, commonly referred to as an avatar
- a (video of a) virtual human that can be synthesised with generative AI methods based on real human video/image data (see the work of Stoll et al. (2020)).

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

Text to AMR

- Abstract meaning representation (Banarescu et al., 2013)
- Represents (“extracts”) meaning from a given sentence
- Lexicon is in “English” regardless of input language
- We developed EN/NL/ES → AMR multilingual (91.8% acc) and EN → AMR monolingual (93.6% acc) neural model (based on mBART)

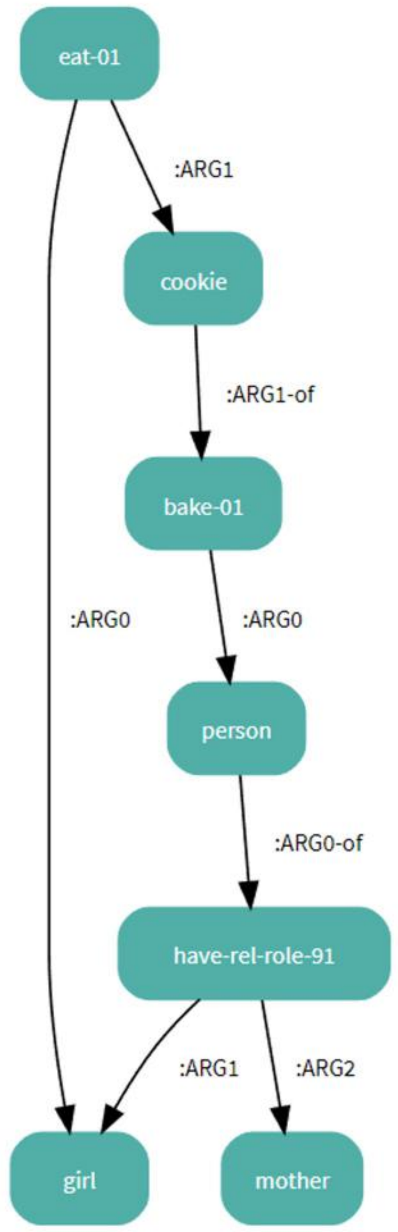
Ex. **The girl eats the cookies that her mother baked**

1. **Linearized AMR form**

```
eat-01 girl cookie bake-01 person have-rel-role-91 mother
```

2. **Concepts only:**

```
eat girl cookie bake mother
```



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

[1]Demo: https://huggingface.co/spaces/BramVanroy/text-to-amr

Look up concepts in modified SignBanks

We modified the NGT/VGT SignBanks to include English “translations” so we can do reverse look-ups (English word → gloss)

- Multilingual WordNet
- ChatGPT: “context-sensitive” translations
- Similarity filtering with LABSE vectors

So continuing with English concepts extracted from AMR… eat girl cookie bake mother

3. Reverse look-up the English AMR concepts → gloss (ex. is VGT)

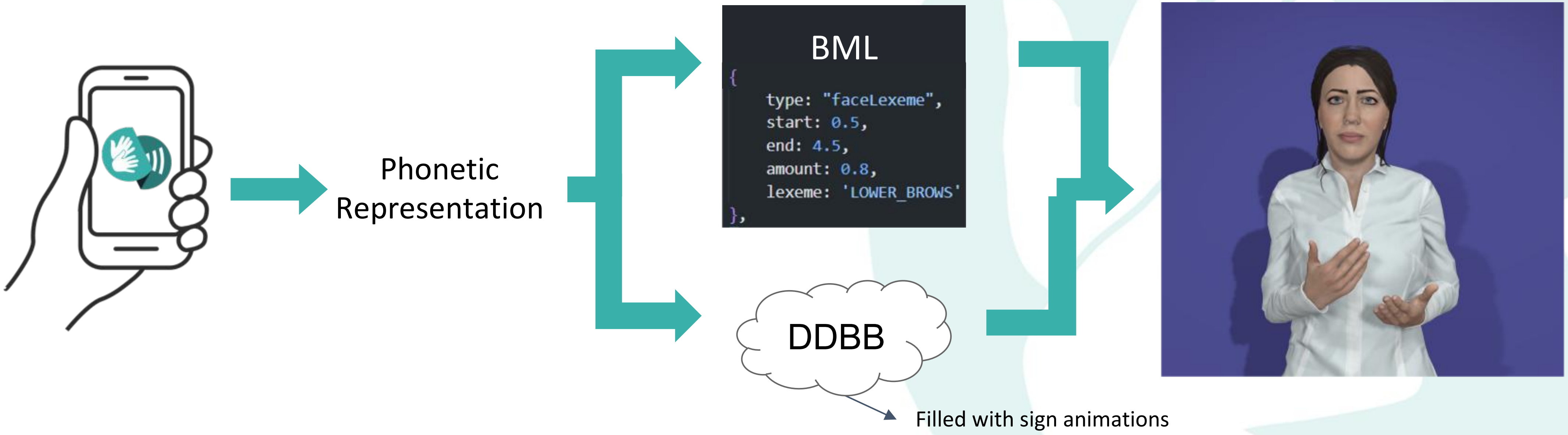
   ETEN-A MEISJE-B KOEK-C BAKKEN-A MOEDER-A

4. Remove regional identifier

   ETEN MEISJE KOEK BAKKEN MOEDER

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

SignON Synthesis Pipeline



System 1 - Data Generation T5.2

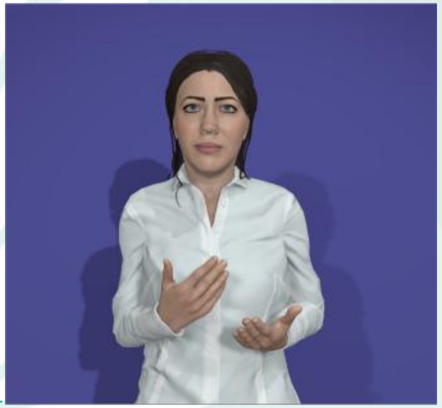
- Capture and edit sign animations
- Obtain sign language animations

System 2 - Data Realisation T5.1, T5.4, T5.5

- Render scene to be used in the mobile app
- 3D engine to animate a virtual avatar

Current Overview

Phonetic Representation → BML →



Top-to-bottom Synthesis Strategy

- Top: sign animations

VGT

NGT

Any SL

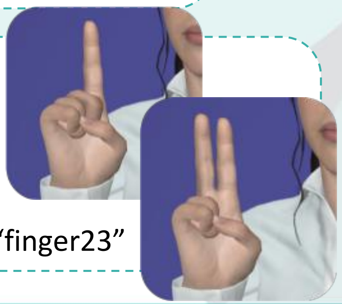
This encoding is **language independent**. Meaning it can be used for any SL if the data exist

- Bottom: spatial behaviours

Extensive work is being done to support all possible base behaviours → High level of control

Handshape: “finger2”

Handshape: “finger23”



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

Focus Tasks

Top-to-bottom Synthesis Strategy

❖ Top: sign animations

Automatic estimation of NMFs

Edition of NMF behaviours

Exportability of final animation

MFs edition with IK solvers

❖ Bottom: spatial behaviours

Substantial work on supporting available datasets in HamNoSys and SiGML encoding.

\+ Contribution to reduce the scarcity of data in sign language.

\+ Support of previous sign projects.

\+ Support of sign language research groups.

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

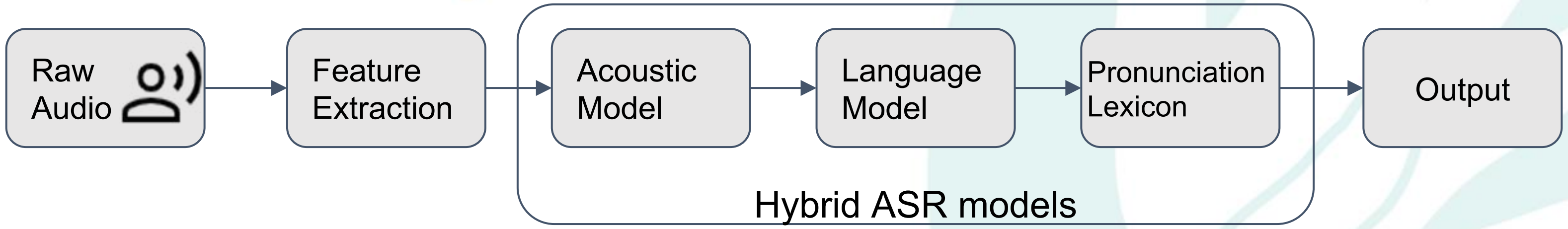
SignON is not only about sign to sign translation

Objective 3: Automated Recognition and Understanding of Signed and Spoken/Verbal Language Input

Automatic Speech Recognition

Current ASR Web Service

KALDI



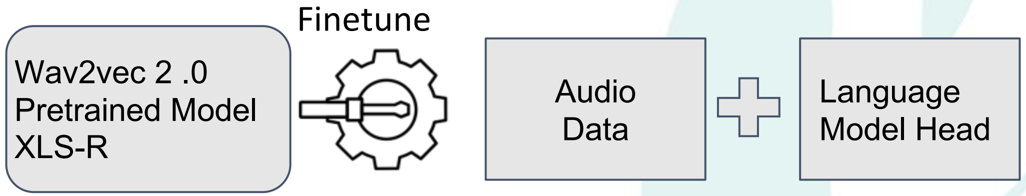
Supported Languages for Typical Speech

- English
- Spanish
- Dutch
- Irish

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

What’s Next: End-to-end Speech Recognition Models

New ASR Web Service based on `Wav2vec 2.0` and `Whisper`

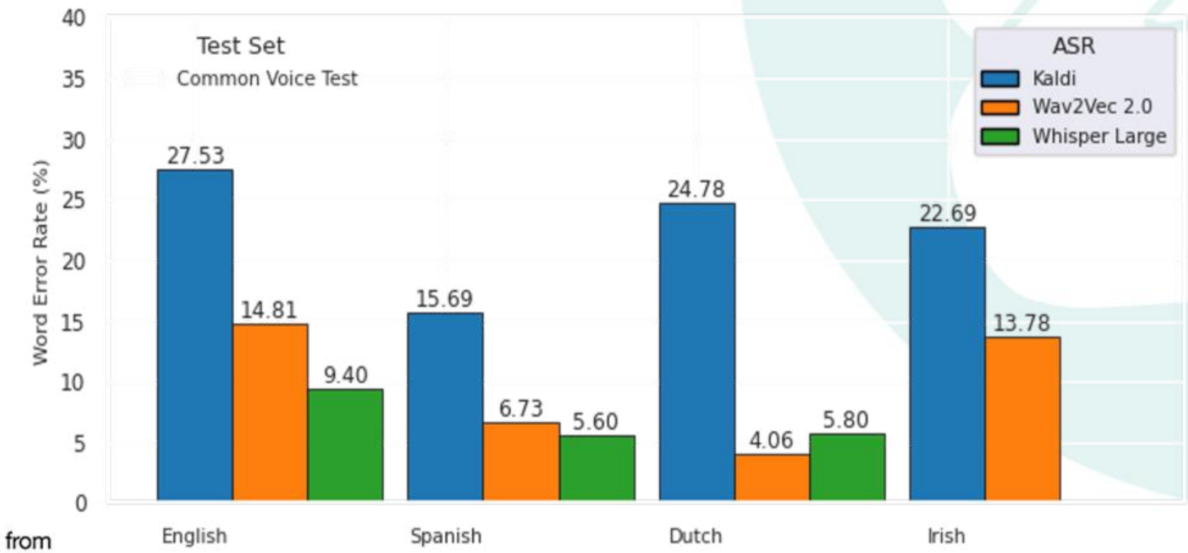


Supported languages for typical speech

- English
- Spanish
- Dutch
- Irish

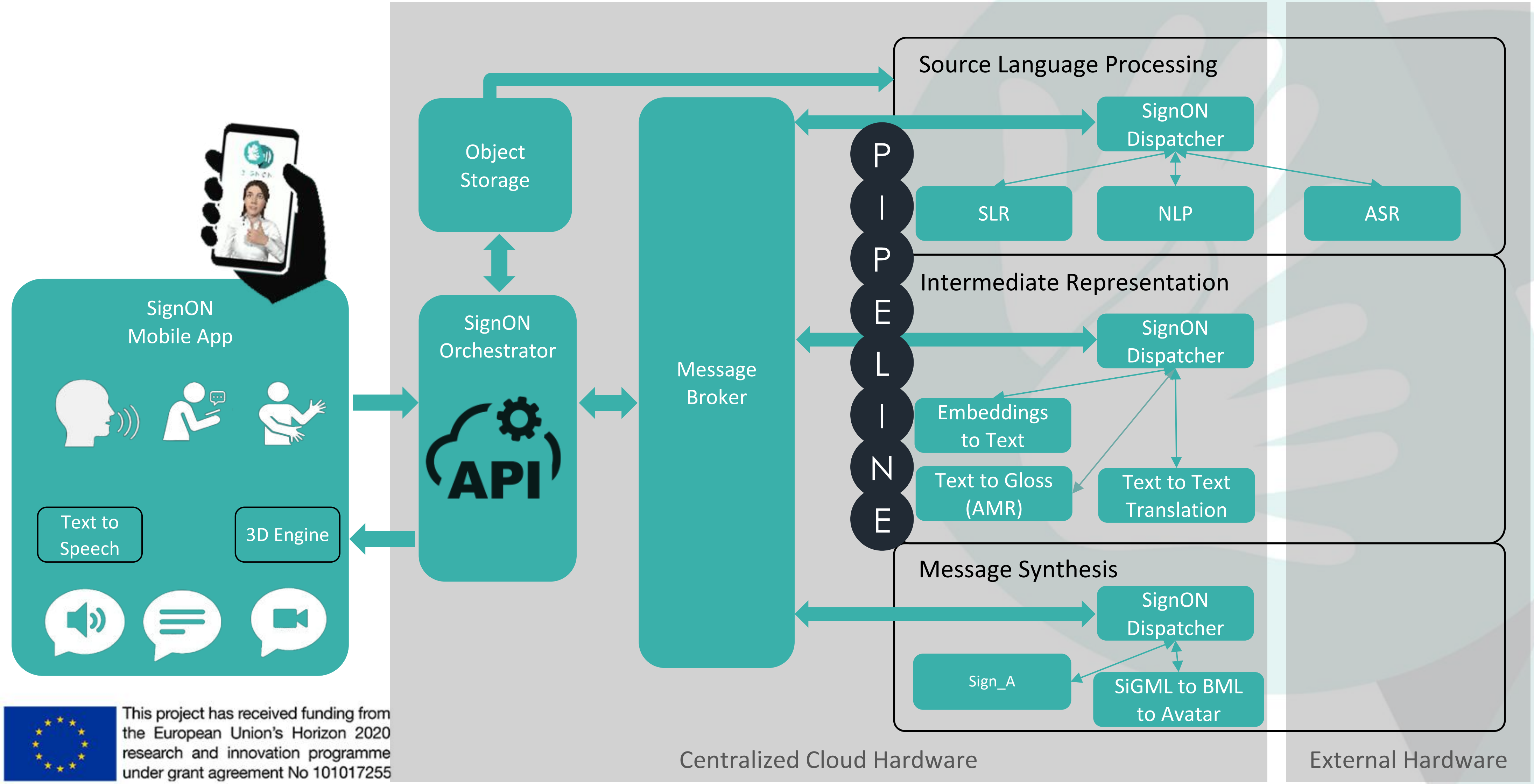
Atypical Speech ASR

Finetuning with use-case recordings



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

Cloud Platform Architecture



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

Thank you!

SIGNON

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