



EASIER & NEOLOGISMS

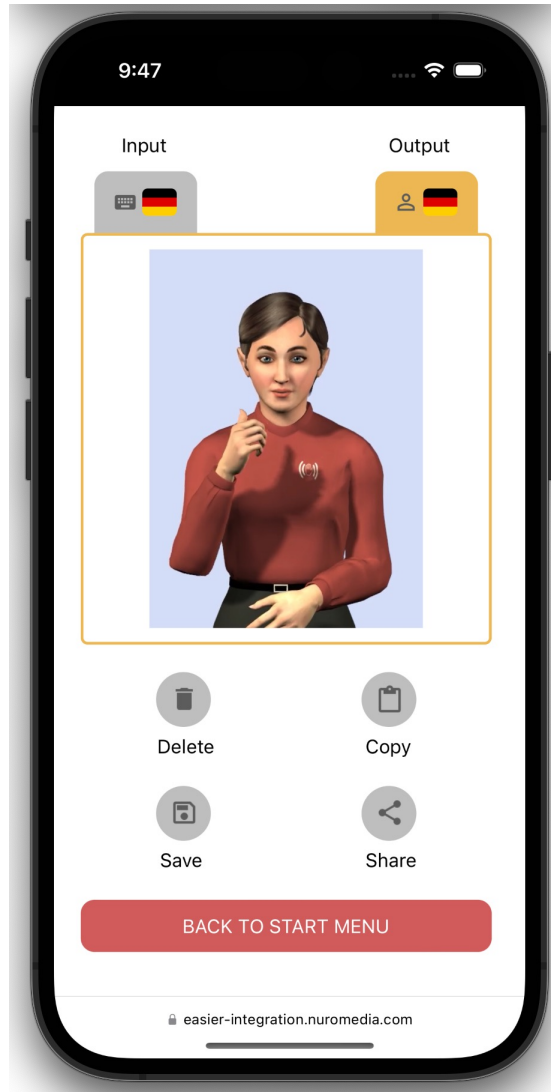
Thomas Hanke
Universität Hamburg

Athens
10.02.2023

Meet the EASIER Consortium



- Translation between sign languages and spoken languages (speech or writing)
- Multiple languages
- Automatic (near-realtime) and semi-automatic (human post-editing)
- Mobile app demonstrating the automatic route
- Goals shared with the SignOn project (both funded within EU ICT-57)



- Speech-to-text
→ State-of-the-art machine translation
→ Avatar presenting signed output
- Robust data-driven video recognition →
State-of-the-art machine translation →
Output in text

- Producing comprehensible and easy-to-read signing is a must:
 - Producing a signed utterance sign after sign simply is not enough!
 - Communication is more than content-passing.

- Quality improvements in avatar technology beyond producing signs one by one: Prosody, affect



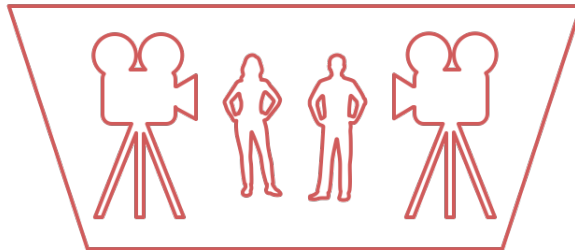
- Quality improvements in avatar technology beyond producing signs one by one: Prosody, affect
- End users in the loop: Develop measures for “ease of listening” once comprehensibility is achievable.
- Explore innovative ways for post-editing not only for expert translators, but accessible to community members.



- Sign language resources needed to train machine learning models are smaller than available spoken language data **by several orders of magnitude.**

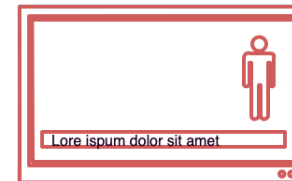
- We therefore combine linguistic corpora and interpreted broadcast footage.

Linguistic corpora



High quality
Variety of elicitation tasks
Source: Sign Language
Semi-spontaneous language production
Rich linguistic annotation + translations

Broadcasting data



Large quantity
News domain
Source: spoken language
Interpreting under time pressure
Subtitles/Captions

Language	Number of sentence pairs in corpus*	Number of sentence pairs in broadcast data**
DGS	64 000	1 130 000
BSL	6 000	1 150 000
Std corpus for English-German	150 000 000	150 000 000

* Public DGS Corpus, BSL Corpus

** Broadcast data prepared for processing within EASIER

- We have a multitude of sign/spoken language pairs in Europe, ranging from some data to practically none.
- We try to create synergies by treating pairs in parallel.

- So why is EASIER interested in neologisms?
 - The idea that for each content word in one language you find a corresponding word in the other language is an extreme simplification of the game, esp. wrt sign language.
 - If constructed action, classifiers and such come to your mind: Yes, right, but the problems are already there on the lexical level!

- So why is EASIER interested in neologisms?
 - With our training data not only being orders of magnitude too small, but also being old, the data may well not contain counterparts in the other language.
 - And then not all lexical items do have counterparts in the other language: Lexical gaps occur within any language pair you can think of.
 - Sometimes neologisms close gaps. With so many gaps around, we better understand how they work.

- And our avatar should be prepared to produce signs that have not been around when it was trained.
 - See today's poster/demo by Wolfe et al.



THANKS



EASIER project is funded by the EU's Horizon2020 Programme under Grant Agreement number 101016982



[PROJECT-EASIER.EU](https://project-easier.eu)



[@EASIERPROJECT](https://twitter.com/EASIERPROJECT)