easier

intElligent Automatic
SIgn languagE tRanslation

D4.3 EASIER FINAL TRANSLATION SYSTEMS V2

Revision: v1.2

Work Package	WP4
Task	T4.2, T4.3
Due date	31/07/2023
Submission date	31/07/2023
Deliverable lead	University of Zurich
Version	1.2
Authors	Mathias Müller (University of Zurich – UZH), Annette Rios (University of Zurich – UZH), Amit Moryossef (University of Zurich – UZH), Sarah Ebling (University of Zurich – UZH)
Reviewers	Eleni Efthimiou, Rosalee Wolfe (ATHENA), Thomas Hanke (UHH)

Abstract	This deliverable describes the final version of the spoken-to-sign and sign-to-spoken translation models developed in Task T4.2 and the final version of the spoken-to-spoken translation models developed in Task T4.3.
Keywords	sign language translation, machine translation, gender bias

Grant Agreement No.: 101016982
Call: H2020-ICT-2020-2
Topic: ICT-57-2020
Type of action: RIA

Document Revision History

Version	Date	Description of change	List of contributors
V1.0	24/07/2023	First draft	Mathias Müller, Annette Rios (UZH)
V1.1	25/07/2023	Internal Review 1	Thomas Hanke (UHH)
V1.2	27/07/2023	Internal Review 2	Eleni Efthimiou, Rosalee Wolfe (ATHENA)

DISCLAIMER

The information, documentation and figures available in this deliverable are written by the “Intelligent Automatic Sign Language Translation” (EASIER) project’s consortium under EC grant agreement 101016982 and do not necessarily reflect the views of the European Commission.

The European Commission is not liable for any use that may be made of the information contained herein.

COPYRIGHT NOTICE

Project co-funded by the European Commission in the H2020 Programme		
Nature of the deliverable	R	
Dissemination Level		
PU	Public, fully open, e. g., web	✓
CL	Classified, information as referred to in Commission Decision 2001/844/EC	
CO	Confidential to EASIER project and Commission Services	

* R: Document, report (excluding the periodic and final reports)
DEM: Demonstrator, pilot, prototype, plan designs
DEC: Websites, patents filing, press & media actions, videos, etc.
OTHER: Software, technical diagram, etc

EXECUTIVE SUMMARY

This deliverable reports on the machine translation systems developed in the EASIER project. Our earlier deliverable (D4.2) reported on the first version of the translation systems, the current deliverable describes the final systems that are delivered.

The deliverable firstly gives an overview of the parallel corpora curated by EASIER that serve as the training and evaluation data for machine translation. Secondly, we report on experiments on sign language machine translation on five language pairs, with conclusions and recommendations about which exact system and procedure perform best. The best-performing systems are then delivered for use in EASIER and for the general public.

Finally, we also describe experiments on spoken language machine translation, focusing specifically on gender bias considerations.

CONTENTS

Funded by the Horizon 2020
Framework Programme of the European Union

Funded by the Horizon 2020
Framework Programme of the European Union

LIST OF FIGURES

Funded by the Horizon 2020
Framework Programme of the European Union

LIST OF TABLES

Funded by the Horizon 2020
Framework Programme of the European Union

ABBREVIATIONS

EMSL	European Meta-Sign Language
MT	machine translation
NMT	neural machine translation
SQM	Scalar Quality Metric
CNN	Convolutional Neural Network
EMSL	European Meta Sign Language
TAG	Tree Adjoining Grammar
SLVA	Sign Language Video Anonymisation
GCP	Google Cloud Platform
DE	German
EN	English
IT	Italian
FR	French
NLP	Natural Language Processing
ASL	American Sign Language
BSL	British Sign Language
DGS	German Sign Language / Deutsche Gebärdensprache
DSGS	Swiss-German Sign Language / Deutschschweizer Gebärdensprache
LIS	Italian Sign Language / Lingua Italiana dei Segni
LSF	French Sign Language / Langue des Signes Française

Funded by the Horizon 2020
Framework Programme of the European Union

1 INTRODUCTION

This deliverable reports on machine translation (MT) research conducted in the EASIER project. The overall goal of this research is to improve the quality of automatic sign language translation technology. More specifically, the purpose of our experiments was to demonstrate which exact techniques lead to the best translation quality and should therefore be used in a production system. Overall we conducted a series of comparative experiments, comparing many ways of combining datasets, existing and entirely novel technologies. The experiments cover five language pairs: (British Sign Language (BSL)↔English (EN), German Sign Language (DGS)↔German (DE), Swiss-German Sign Language (DSGS)↔DE, Italian Sign Language (LIS)↔Italian (IT) and French Sign Language (LSF)↔French (FR)).

All our efforts involving sign language are detailed in Chapter 2. We pursued fundamentally different approaches depending on the translation direction, that is why spoken-to-signed and signed-to-spoken translation are reported on in separate chapters. For signed-to-spoken translation systems (described in Section 2.1), our focus is on determining the best representation for sign language data to be used in a translation system. For spoken-to-signed translation (described in Section 2.2), our focus is to establish the first publicly available baseline systems, together with the first open-source implementations.

For some spoken-to-signed systems that generate sign language utterances as an output, we also report on a preliminary human evaluation.

Finally, we conducted experiments on spoken-to-spoken machine translation to complement our sign language experiments (described in Chapter 3), focusing on making conventional spoken language MT available for EASIER and experiments on gender bias.

2 SIGN LANGUAGE TRANSLATION (TASK T4.2)

We report separately on signed-to-spoken translation (Section 2.1) and spoken-to-signed translation (Section 2.2). This is because our technical approach is different for each translation direction.

2.1 SIGNED-TO-SPOKEN TRANSLATION

Overall, our goal for signed-to-spoken systems is translating from a signed input video to a spoken output text, by any means that are technically feasible. We are agnostic to the exact technologies used, as long as the entire process is fully automatic – this is reflected in our experiments which are comparisons of many different approaches (Section 2.1.1).

For the experiments we rely on broadcast data collected and curated by EASIER partners (Section 2.1.2). The technology we have developed is entirely novel in many cases. Since the training of such systems is not tried-and-tested, we began a series of exploratory experiments on the DSGS-DE language pair (Section 2.1.3). These preliminary experiments inform the exact parameters and decisions for the main experiments on five language pairs (Section 2.1.4).

The code to reproduce all of our experiments is publicly available here: `https://github.com/ZurichNLP/easier-continuous-translation`.

2.1.1 Overview of technical approaches

Existing translation systems for signed-to-spoken translation can be broadly categorized by how sign language is represented in each system. Well-known paradigms (illustrated in Figure 2.2) are the following:

- **Gloss translation systems:** Sign language data is represented as *glosses*, which are semantic labels borrowed from a related spoken language, often one gloss for each individual sign. Representing sign language as glosses reduces the machine translation problem to a conventional text-to-text translation problem and therefore, all previous research in machine translation can be applied most readily. However, glosses as a representation have distinct shortcomings, and few resources have gloss annotations (Müller et al., 2023).
- **Pose translation systems:** Sign language data is represented as a graph of body keypoints in 2-dimensional or 3-dimensional space. Extracting such keypoints, *pose estimation*, is a fully automatic process, well-known pose estimation systems are OpenPose (Cao et al., 2021)[1] and MediaPipe Holistic (Lugaresi et al., 2019)[2]. The exact set of keypoints depends on the pose estimation system. Poses are more lightweight than original video frames, more tailored towards the task at hand (since they focus on movement,

[1] `https://github.com/CMU-Perceptual-Computing-Lab/openpose`

[2] `https://ai.googleblog.com/2020/12/mediapipe-holistic-simultaneous-face.html`

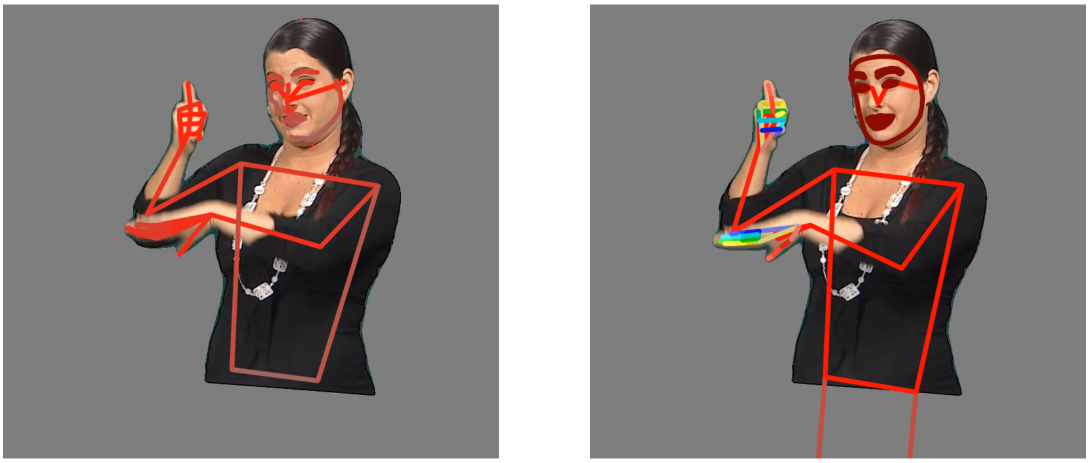


Figure 2.1: *Examples of the output of pose estimation systems overlaid over the original video frames. Left: OpenPose, right: MediaPipe Holistic.*

rather than other information in a video such as the background) and potentially more anonymous than original videos. However, recent experiments have shown that pose information does not represent sign language as well as more generic video feature extractors (Moryossef et al., 2021a; Müller et al., 2022; Tarrés et al., 2023).

- **Numerical feature extraction:** Finally, translation systems can encode (on the input side) or generate (on the output side) sign language represented as arbitrary numerical data that implicitly encodes meaning. In the context of machine learning and Natural Language Processing (NLP), such numerical representations are often called *embeddings* or learned *features*. The most well-known mechanism to extract such features is to train a Convolutional Neural Network (CNN) to solve a preliminary task (other than the translation task) such as single-sign recognition in videos with continuous signing. As a by-product of learning to solve this preliminary task, the CNN will learn to transform high-dimensional input videos into a smaller numerical representation.

EMSL representations Besides the standard ways of representing sign language data explained above, EASIER also developed custom representations called the European Meta Sign Language (EMSL). EMSL is explained in more detail in **Deliverable 3.3** and **Deliverable 3.4**. Here we explain only what is necessary for the purpose of this deliverable.

There are two major versions of EMSL, referred to as V1.0 and V2.0. The fundamental difference between them is that EMSL V1.0 is a numerical representation of sign language, while EMSL V2.0 is a gloss-based representation, consisting of text strings. See Figure 2.3 for a high-level comparison. There are further, minor variants of each major EMSL version. More specifically,

- **EMSL v1.0a:** a latent representation obtained by contrastive learning. Now discontinued and not used for our translation experiments.
- **EMSL v1.0b:** a numerical, intermediate output from learning a continuous sign language recognition (or *spotting*) task. The general approach is to use as a starting point a 3D-

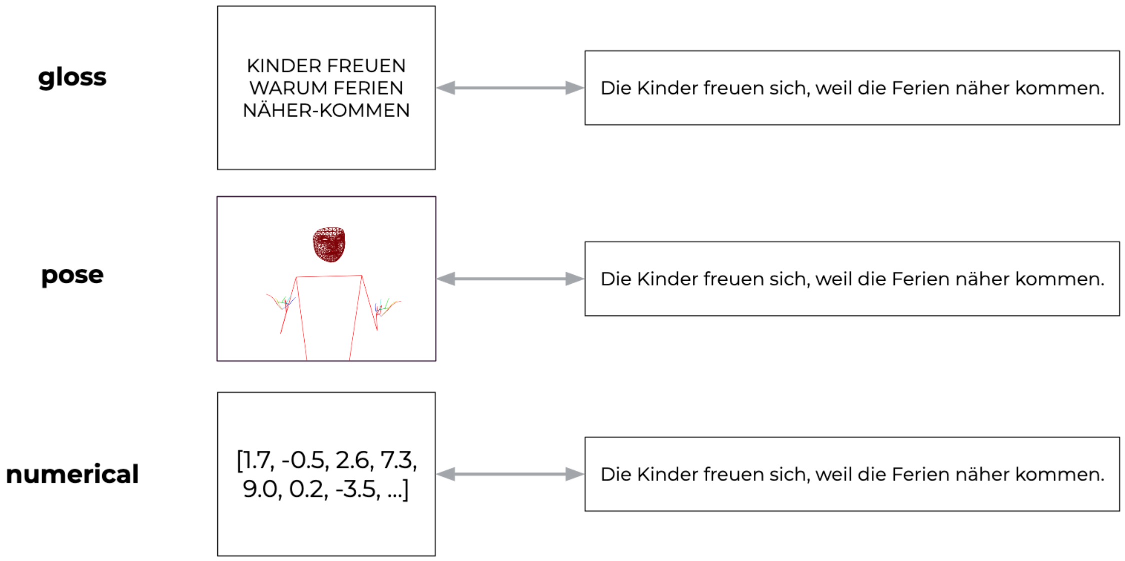

gloss
KINDER FREUEN WARUM FERIEN NÄHER-KOMMEN
Die Kinder freuen sich, weil die Ferien näher kommen.
pose
Die Kinder freuen sich, weil die Ferien näher kommen.
numerical
[1.7, -0.5, 2.6, 7.3, 9.0, 0.2, -3.5, ...]
Die Kinder freuen sich, weil die Ferien näher kommen.

Figure 2.2: *Types of sign language translation systems, exemplified with an example in Swiss German Sign Language (DSGS, left) and German (right). The main differences between system types is how sign language is represented. The illustration shows double arrows to indicate that in principle, is possible translation in both directions.*

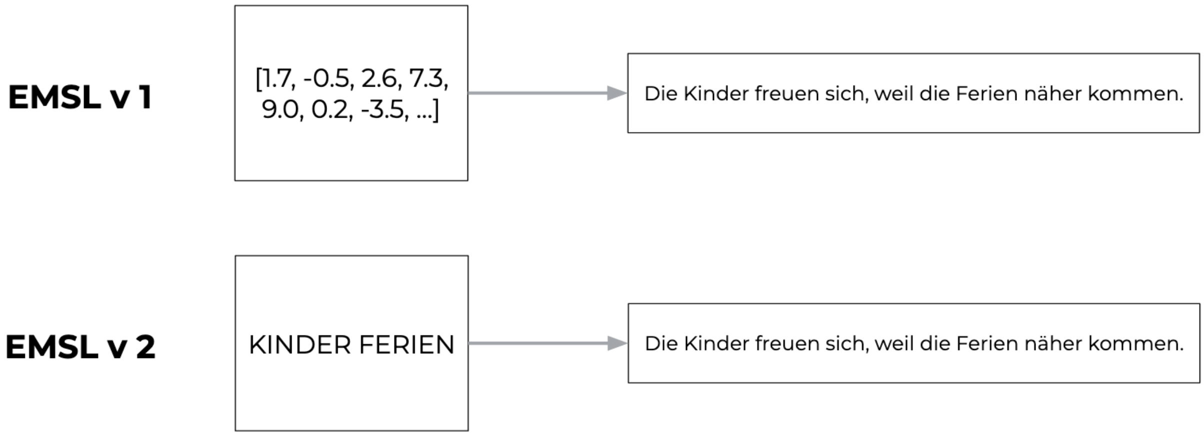

EMSL v 1
[1.7, -0.5, 2.6, 7.3, 9.0, 0.2, -3.5, ...]
Die Kinder freuen sich, weil die Ferien näher kommen.
EMSL v 2
KINDER FERIEN
Die Kinder freuen sich, weil die Ferien näher kommen.

Figure 2.3: *Comparison of systems based on EMSL versions 1.0 and 2.0. A fundamental difference between them is that EMSL V1.0 is a numerical representation of sign language, while EMSL V2.0 is a gloss-based representation, consisting of text strings. This figure is an extension of the basic system types introduced in Figure 2.2.*

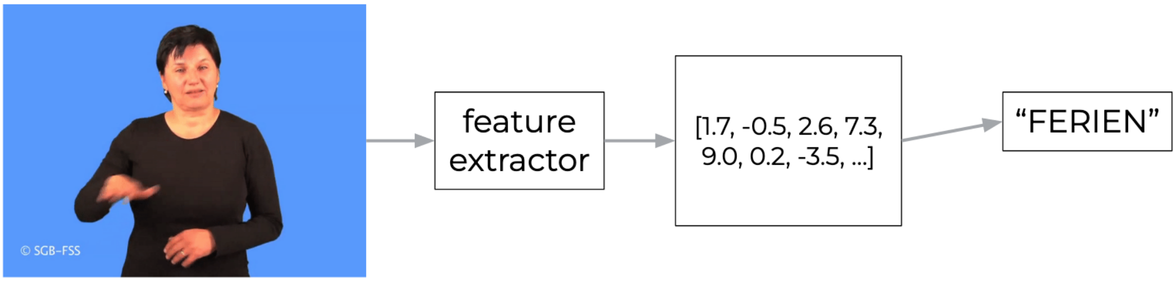


Figure 2.4: *Informal illustration of EMSL version 1.0b. A system is trained to recognize glosses (sign classes) in continuous videos. In a first step, numerical features are extracted from the original video, for instance with a CNN. For gloss recognition, the extracted features are an intermediate output that is then used to predict actual glosses. Once this model is trained, the intermediate extracted features are used as EMSL v1.0b and the second step (of predicting classes) is omitted.*

CNN model pre-trained for general action recognition (Carreira and Zisserman, 2017) and fine-tune it on sign language data. See Figure 2.4.

- **EMSL v2.0a:** the final output of a gloss recognition (or *spotting*) system. The system takes existing videos and subtitles as an input and generates a sequence of glosses as an output. For all the spoken language words of an existing subtitle, the system predicts where these words occur in the video. Since this approach relies on subtitles to constrain the search space, it is a theoretical *oracle experiment*, and cannot be used in practice. In practice, we assume that subtitles are not available while generating a new translation with a trained translation system.
- **EMSL v2.0b:** same as v2.0a except that the system predicts from the full set of all known glosses (classes of signs), hence a more realistic and practical setting. EMSL v2.0b in turn comes in different versions, depending on two important hyperparameters: what data the spotter was trained on (Public DGS Corpus (Hanke et al., 2020) or BOBSL (Albanie et al., 2021)) and a probability threshold for detecting glosses in the video. See Figure 2.5.

2.1.2 Data

Our translation systems use as training data the resources collected by EASIER partners in an earlier phase of the project. The focus of that data collection effort was on news broadcast material, harvested from the websites of public broadcasters in the respective countries. See an overview of the available parallel data in Table 2.1. For BSL↔EN, EASIER did not collect data on its own and instead relies on data previously collected by Albanie et al. (2021). The data collection and curation is described in more detail in **Deliverable 4.1**.

Sentence segmentation Subtitle units from news broadcast material do not necessarily correspond to full sentences. Entire-sentence units are required for machine translation experiments. We therefore used automatic sentence segmentation methods to re-distribute the text

Funded by the Horizon 2020 Framework Programme of the European Union

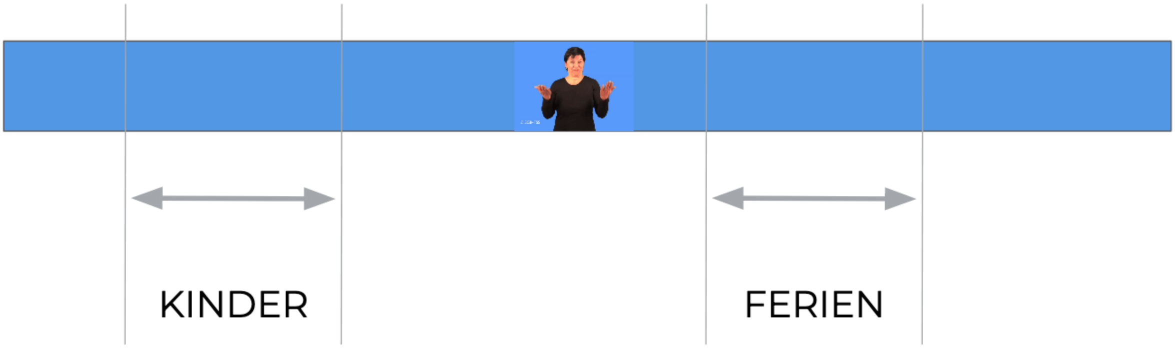


Figure 2.5: *Informal illustration of EMSL version 2.0b. A system is trained to recognize glosses (sign classes) in continuous videos. At inference time, the model takes a video as an input and predicts where in the video specific signs occur. The set of signs that can be predicted depends on the data that this gloss recognition model was trained on (DGS or BSL data).*

language pair	total duration in hours	number of videos with subtitles available
BSL↔EN	1462	-
DGS↔DE	2171	1928
DSGS↔DE	4647	1928
LSF↔FR	918	798
LIS↔IT	111	230

Table 2.1: *News broadcast resources collected by EASIER in five language pairs. Table shows the parallel corpora available after preprocessing. For BSL, the existing resource BOBSL was used. Numbers are taken from Deliverable 4.1.*

Funded by the Horizon 2020
Framework Programme of the European Union

Example 1	
Original subtitle	After automatic segmentation
81 00:05:22,607 -> 00:05:24,687 Die Jury war beeindruckt 82 00:05:24,687 -> 00:05:28,127 und begeistert von dieser gehörlosen Frau.	48 00:05:22,607 -> 00:05:28,127 Die Jury war beeindruckt und begeistert von dieser gehörlosen Frau.
Example 2	
Original subtitle	After automatic segmentation
7 00:00:24,708 -> 00:00:27,268 Die Invalidenversicherung Region Bern startete 8 00:00:27,268 -> 00:00:29,860 dieses Pilotprojekt und will herausfinden, ob man es 9 00:00:29,860 -> 00:00:33,460 zukünftig umsetzen kann. Es geht um die Umsetzung	4 00:00:24,708 -> 00:00:31,720 Die Invalidenversicherung Region Bern startete dieses Pilotprojekt und will herausfinden, ob man es zukünftig umsetzen kann. 5 00:00:31,720 -> 00:00:34,502 Es geht um die Umsetzung (...)

Table 2.2: *Examples of automatic sentence segmentation for German subtitles. The subtitles are formatted as SRT, a common subtitle format.*

of the original subtitles so that each subtitle corresponds to one well-formed sentence exactly. This process is illustrated in Table 2.2.

Alignment shift A further peculiarity of news broadcast data is that the signing is mostly produced by live interpretation. Moreover, the subtitles are also produced live, or are pre-produced. As a consequence, while both the subtitles and the signing are based on the original speech (audio), due to the live subtitling and live interpreting scenario, a temporal offset between audio and subtitles as well as audio and signing is inevitable. This offset or "alignment shift" is visualized in Figure 2.6.

The presence of such alignment shifts means that essentially, all of our training data should be regarded as *comparable* parallel corpora, with an uncertain correspondence between sign language and spoken language utterances (Etchegoyhen and Gete, 2020). This means not only is our training data not immediately useful for training MT systems, but also that such data can hardly be used as evaluation data.

Manual alignment and evaluation data With the help of other EASIER partners we manually corrected a fraction of all available broadcast resources, to use them as more reliable

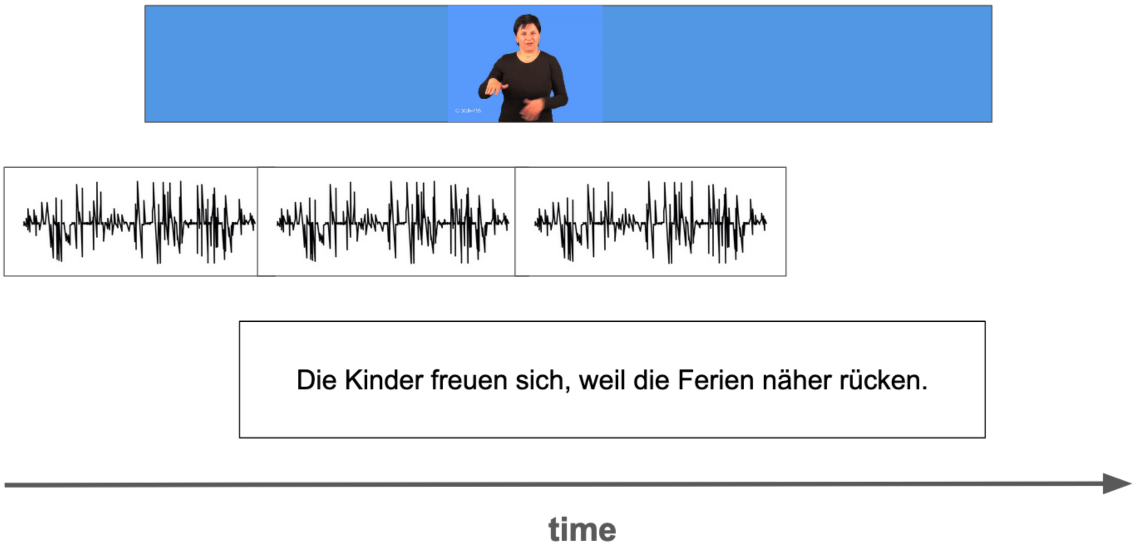


Figure 2.6: *Illustration of alignment shift in sign language corpora. From top to bottom: a sign language video, an audio track with speech, a spoken language subtitle in German. Information in these three modalities do not start and end at the same time, adjusting their start and end times is referred to as alignment.*

training data (most of the manually corrected data) or as evaluation data (one episode for each language pair). At the time of writing such manual corrections were done as follows:

- **DSGS**: 31 episodes were manually corrected as part of the WMT 2022 shared task on sign language translation (Müller et al., 2022)
- **LIS**: 20 episodes were corrected by SWISS TXT
- **LSF**: 20 episodes were corrected by Interpretis
- (**BSL**: manually corrected data already existed and was produced by Albanie et al. (2021), independent of EASIER)

Currently, no human-corrected training or evaluation data exists for DGS. Overall, for our experiments, we adopt the convention of referring to our original, un-aligned training data as **comparable** data and the manually corrected data as **parallel** data.

2.1.3 Preliminary experiments with DSGS data

We ran preliminary experiments with DSGS→DE, as this was the first language pair where larger-scale news data and a stable benchmark were available to us. We explored different system types, ways of combining training data and different training techniques.

Funded by the Horizon 2020
Framework Programme of the European Union

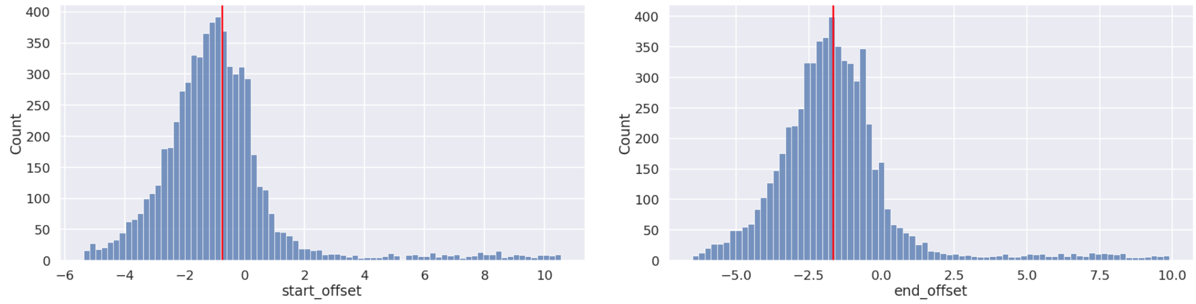


Figure 2.7: *Statistics of alignment shift on the SRF news broadcast data that was corrected manually. start_offset=difference in seconds that a manual annotator has shifted the beginning of a subtitle, end_offset=difference in seconds that a manual annotator has shifted the end of a subtitle. Negative values mean that subtitle times were shifted to a later time during manual correction. Red line shows the median offsets.*

2.1.3.1 Analysis and prediction of alignment shift

Analysis of manually aligned data Since our training data exhibits considerable alignment shift (see Section 2.1.2), we analyzed the DSGS data that was manually corrected to investigate whether the shift can be undone with a simple, statistical method. Figure 2.7 summarizes by how much subtitle times were shifted by our manual annotators during correction. Our analysis shows that the distribution is likely too wide to be captured by a single offset value like a median. This observation is in line with what was observed earlier by Albanie et al. (2021) on the BOBSL data that was manually aligned, and also confirmed empirically by our experiments in Table 2.4 (explained below).

Offset prediction Given that simply adding or subtracting a median offset value is not feasible, we made several attempts to learn an offset prediction model, ranging from a simple regression model which does not take into account the subtitle text to a pre-trained Transformer to encode the subtitle text and a regression head to predict the offset. In the end, all of these attempts were unsuccessful, since none of these prediction methods outperformed the median predictor.

2.1.3.2 Machine translation experiments

We now turn to preliminary machine translation experiments on the DSGS↔ DE training data. Overall, we explore different ways of putting together training data (e.g. only the parallel data, or also adding the comparable data), and different representations for sign language (pose estimation variants, EMSL variants).

Core model For all experiments we use Sockeye (Hieber et al., 2022) as the underlying sequence-to-sequence toolkit, which is based on Pytorch (Paszke et al., 2019). For numerical representations of sign language (see Section 2.1.1) we adapted Sockeye so that it supports

BLEU	`BLEU+case.mixed+numrefs.1+smooth.exp+tok.13a+version.1.4.14`
CHRF	`chrF2+numchars.6+space.false+version.1.4.14`

Table 2.3: *SacreBLEU signatures for evaluation metrics.*

encoding or decoding continuous vectors instead of discrete sequences of tokens. Our implementation can read a sequence of feature vectors and convert them to the model size with a simple learned projection. Apart from this modification, all models are standard Transformer models (Vaswani et al., 2017).

Preprocessing The spoken language subtitles are segmented automatically (see Section 2.1.2). As an experiment variant, for our comparable training data, we shift all subtitle times by the median offset taken from our statistical analysis (see Section 2.1.3.1). The subtitle text is then segmented by a trained Sentencepiece model (Kudo, 2018) with a vocabulary size of 1000.

If pose estimation is used, we test different ways of normalizing pose data (standardizing, for instance, the shoulder width of each frame). For pose or EMSL systems we test forcing a common framerate for all continuous sequences.

Automatic evaluation We perform an automatic evaluation of translation quality. We measure translation quality with BLEU (Papineni et al., 2002) and CHRF (Popović, 2016), computed with the tool SacreBLEU (Post, 2018). Both metrics are widely used in MT research, and SacreBLEU is the recommended tool to compute them. See Table 2.3 for all SacreBLEU signatures. We note that many recent neural metrics, such as COMET (Rei et al., 2020), are not applicable in our case because the source languages (sign languages) are not supported.

2.1.3.3 Results and conclusions

Exploration of EMSL v2.0b variants EMSL v2.0b exists in 12 different variants, depending on what data the spotter was trained on (Public DGS Corpus or BOBSL) and a probability threshold for detecting glosses in the video (see Section 2.1.1). We compare all of these variants, and the outcome of these experiments are summarized in Figure 2.8.

We conclude from our experiments that **using an I3D threshold of 0.5 leads to the highest translation quality** (as measured on the development set, not the test set). The exact I3D model used (based on DGS or BSL glosses, or both combined) is less important as a hyperparameter, since there is no version which is clearly superior. In subsequent results we will show only the one EMSL v2.0b system that achieved the best score on the development set.

Wider comparison of data scenarios and feature types All further experiments are summarized in Table 2.4. In the results we refer to different representations of sign language as *feature types*.

Funded by the Horizon 2020
Framework Programme of the European Union

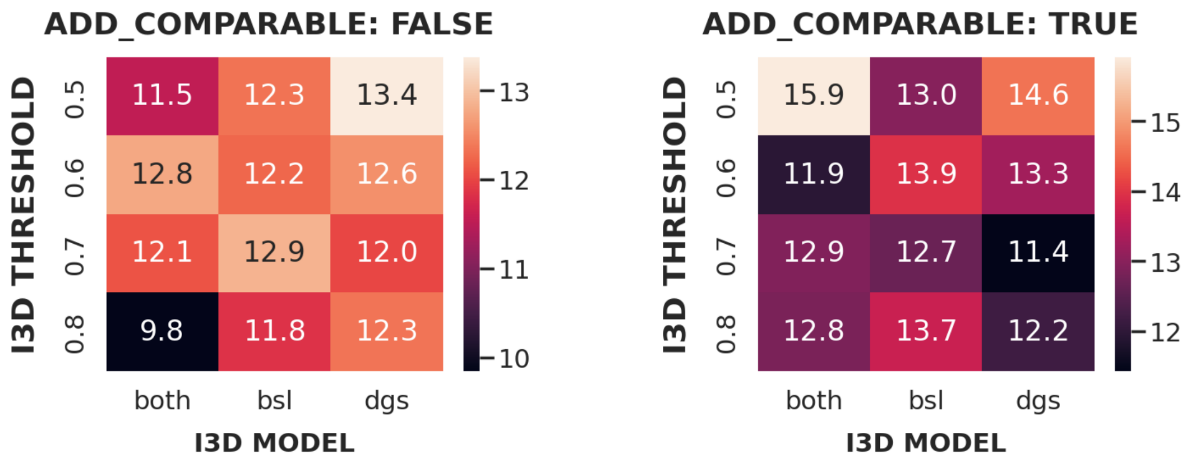


Figure 2.8: *Exploration of different variants of EMSL v2.0b. Numbers are CHRF scores computed on the WMT-SLT22 development set. The I3D threshold and I3D model are hyperparameters of EMSL v2.0b.*

	feature type	training corpora			Translation quality (CHRF)
		parallel	comparable	apply offsets	
(1)	OpenPose	✔	-	-	5.80
(2)	Mediapipe	✔	-	-	7.90
(3)	EMSL v1.0b	✔	-	-	14.00
(4)	EMSL v2.0a*	✔	-	-	17.00
(5)	EMSL v2.0b	✔	-	-	13.00
(6)	EMSL v2.0b	✔	✔	-	15.60
(7)	Mediapipe	✔	✔ (10k)	-	5.57
(8)	Mediapipe	✔	✔ (10k)	✔	4.45
(9)	Mediapipe	✔	✔ (50k)	-	7.00
(10)	WMT-SLT22 winning system	-	-	-	19.50

Table 2.4: *Preliminary experiments on DSGS→DE, showing the translation quality of different signed-to-spoken systems. All experiments are evaluated on the official test set of the WMT-SLT22 shared task (Müller et al., 2022). The WMT-SLT22 winning system is trained on different training data. parallel=EASIER news broadcast data that was manually corrected (identical to SRF training corpus provided by WMT-SLT22), comparable=larger EASIER news broadcast data with alignment shifts, apply offsets=shift subtitle times automatically by an empirical value resulting from an analysis of the manually corrected parallel data, *=EMSL v2.0a is an oracle experiment*

Comparison of pose estimation systems (rows (1) and (2)) We observe that **Mediapipe poses usually outperform OpenPose poses as a representation for sign language**. An additional reason to prefer Mediapipe over OpenPose is that the former runs in real-time on consumer devices, making it more likely that such a system could be used in a handheld application. Regarding the learning procedure of the underlying neural network, we observe that for both pose types, the learning trajectory is uncharacteristic for NMT models. For instance, the perplexity on the training set does not continuously and smoothly decrease over the course of learning. This may indicate that poses need more aggressive normalization.

Promising performance of EMSL v1.0b (rows (1) to (3)) EMSL v1.0b outperforms pose translation systems by a large margin (a CHRF score of 14 versus 7). We conclude that **EMSL v1.0b is a promising representation for sign language machine translation, and more accurate than pose estimation systems**. This finding is in line with earlier research, with most authors concluding that learned vision-based feature extractors outperform current pose estimation systems (Moryossef et al., 2021a; Müller et al., 2022; Tarrés et al., 2023). Finally, the learning trajectory of EMSL v1.0b systems is closer to our expectations, adding to its potential usefulness.

EMSL v2.0 variants worse than alternatives (rows (3) to (6)) While EMSL v2.0a outperforms v1.0b in terms of translation quality, it is an oracle experiment that has no practical application, i.e. cannot be used in a real translation system. The performance of the best hyperparameter combinations for EMSL v2.0b is comparable to EMSL v1.0b. However, for v2.0b considerable hyperparameter tuning was conducted, while no hyperparameter optimization was done for v1.0b. All systems can be improved in similar ways with hyperparameter optimization. In effect, this means that **EMSL v1.0b is superior to all tested EMSL v2.0 variants**.

Adding comparable training data (rows (7) to (9)) In this part of the experiments, we kept the feature type constant, using Mediapipe in all cases. We varied the exact composition of the training data. Our results show that **naively combining our hand-corrected, parallel data with comparable data (from our larger pool of data containing alignment shifts) degrades performance**. This is the case even if a median offset is applied to the comparable subtitles. We observe a slight trend that using more comparable data improves translation quality, while still underperforming a system that is trained on parallel data only. A crucial conclusion we draw is that **a pre-training (on comparable data) and fine-tuning (on parallel data) scheme is necessary**.

Miscellanous findings (not represented in the table) We also established that pose normalization is necessary and that forcing a common framerate (independent of the feature type) improves the translation quality. Finally, none of our systems was able to outperform the winning system of the WMT 2022 shared task on sign language translation, pointing to the fact that our preprocessing, core implementation or training procedures still have room for improvement.

Funded by the Horizon 2020
Framework Programme of the European Union

	parallel						
	OpenPose	Mediapipe	EMSL v1.0b		EMSL v2.0b		
			DGS	BSL	DGS	BSL	BOTH
BSL↔EN	32563	32563	-	32029	-	8276	-
DGS↔DE	-	-	-	-	-	-	-
DSGS↔DE	6855	6855	6855	6855	6993	6975	7019
LSF↔FR	3422	3422	3422	3422	3594	3604	3618
LIS↔IT	3526	3526	3526	3526	3656	3648	3685

	comparable						
	OpenPose	Mediapipe	EMSL v1.0b		EMSL v2.0b		
			DGS	BSL	DGS	BSL	BOTH
BSL↔EN	1191039	1191039	-	1176385	-	799873	-
DGS↔DE	963073	963073	51316	51316	-	-	-
DSGS↔DE	560304	560304	21819	21819	12705	12650	12837
LSF↔FR	256508	256508	23399	256508	22701	242390	23096
LIS↔IT	45015	45015	18296	18296	18384	18326	18531

Table 2.5: *Training data effectively available for all language pairs. This table shows the number of parallel pairs of spoken language sentence, signed utterance, the statistic that is most immediately useful for sentence-level machine translation experiments.*

2.1.4 Main experiments on five language pairs

If not noted otherwise, all preprocessing, model and training settings are identical to our preliminary experiments (see Section 2.1.3).

We do not have full coverage of all datasets for all feature types and all language pairs. Besides partial coverage, we exclude training examples for reasons such as corrupted video files, subtitle annotations that do not match, framerate issues and so on. As a consequence, our effective available training data is less than suggested in the overall numbers from our data collection. Table 2.5 shows the number of parallel training examples (on sentence level) that are available in reality.

The main results on the full set of five language pairs are shown in Table 2.6. While our main results show CHRF scores, an additional table with BLEU scores is included in Appendix 5.3.

Systems trained on parallel data The performance of systems trained only on manually corrected, parallel data confirms the trends we have observed in our preliminary experiments (Section 2.1.3). EMSL v1.0b, a continuous, numerical feature type, outperforms pose estimates and leads to considerably higher translation quality. The exact variant of EMSL v1.0b (EMSL feature extractor trained on DGS or BSL data) does not make a difference empirically, as the results are comparable.

			DSGS→DE	LSF→FR	LIS→IT	BSL→EN	DGS→DE
	(pre)trained on	finetuned on					
Mediapipe	parallel	-	4.690	8.320	4.570	4.990	-
EMSL v1.0b DGS	parallel	-	**15.460**	13.910	13.140	-	-
EMSL v1.0b BSL	parallel	-	15.370	12.370	13.630	12.050	-
EMSL v2.0b DGS	parallel	-	11.536	9.271	13.185	-	-
	both	-	13.306	11.597	11.521	-	-
EMSL v2.0b BSL	parallel	-	13.340	11.240	14.987	10.571	-
	both	-	13.384	**14.526**	**18.066**	**20.261**	-
EMSL v2.0b BOTH	parallel	-	12.456	9.725	14.634	-	-
	both	-	14.926	14.053	16.142	-	-
Mediapipe	comparable	-	8.160	6.210	4.710	5.700	5.570
EMSL v1.0b DGS	comparable	-	14.290	13.120	14.580	-	**15.620**
EMSL v1.0b BSL	comparable	-	14.680	11.260	15.110	12.630	15.430
Mediapipe	comparable	parallel	6.000	4.670	4.590	5.210	-
EMSL v1.0b DGS	comparable	parallel	14.900	11.780	14.590	-	-
EMSL v1.0b BSL	comparable	parallel	15.340	11.250	14.380	11.440	-

Table 2.6: *Translation quality measured by CHRF on the EASIER manually corrected test data. Best scores are highlighted in bold. For BSL↔EN, EMSL v1.0b DGS does not exist, hence no model was trained. Similarly, parallel data does not exist for DGS↔DE, and no such model was trained*

Systems pretrained on comparable data A further set of systems is trained on comparable data that is potentially of lower quality, but more abundant. Because different amounts of usable comparable data exist, depending on the feature type and language pair (see Table 2.5), we use up to 50000 parallel samples for each system. We trained these systems with the intention of further finetuning them on the parallel data (see below), but they can also be evaluated on their own. When using EMSL v1.0b as the feature type, pretraining on the comparable data alone already leads to translation quality comparable to training on the parallel data.

Final recommendations Learned, continuous representations require less attention to hyperparameter settings and achieve comparable results to the discrete sequences in EMSL v2.0b. Both types of EMSL representations are clearly superior for translation quality than simple pose estimation. Instead of naively combining datasets of varying quality, a more elaborate pre-training and fine-tuning scheme can further improve translation quality. In general, training on more data, even if noisy, is beneficial compared to training only on the very limited amount of high quality data.

2.2 SPOKEN-TO-SIGNED TRANSLATION

The research in this section was published as Moryossef et al. (2023). The text was adapted to fit into the context of this deliverable. Our software is publicly available[3].

[3] https://github.com/ZurichNLP/spoken-to-signed-translation

Funded by the Horizon 2020 Framework Programme of the European Union

2.2.1 Introduction

One of the critical issues in this field is the lack of a reproducible and reliable baseline for sign language translation systems. Without a baseline, it is challenging to measure the progress and effectiveness of new methods and systems. Additionally, the absence of such a baseline makes it difficult for new researchers to enter the field, hampers comparative evaluation, and discourages innovation.

Addressing this gap, this work presents an open-source implementation of a text-to-gloss-to-pose-to-video pipeline approach for sign language translation, extending the work of Stoll et al. (Stoll et al., 2018; Stoll et al., 2020). Our main contribution is the development of an open-source, reproducible baseline that can aid in making sign language translation systems more available and accessible, particularly in resource-limited settings. This open-source approach allows the community to identify issues, work together on improving these systems, and facilitates research into novel techniques and strategies for sign language translation.

Our approach involves three alternatives for text-to-gloss translation, including a lemmatizer, a rule-based word reordering and dropping component, and a neural machine translation (NMT) system. For gloss-to-pose conversion, we use lexicon-acquired data for three signed languages, including Swiss German Sign Language (DSGS), Swiss French Sign Language (LSF-CH), and Swiss Italian Sign Language (LIS-CH). We extract skeletal poses using Mediapipe and apply a series of improvements to the poses, including cropping, concatenation, and smoothing, before applying a smoothing filter.

2.2.2 Background

Sign language translation can be accomplished in various ways. In this section, we focus on the pipeline approach that involves text-to-gloss, gloss-to-pose, and, optionally, pose-to-video techniques. The text-to-gloss technique translates spoken language text into sign language glosses, which are then converted into a sequence of poses by gloss-to-pose techniques, and into a photorealistic video using pose-to-video techniques.

This pipeline offers the benefit of preserving the content of the sentence, while exhibiting a tendency for verbosity and a lower degree of fluency. In this section, we explore each of the pipeline components comprehensively and examine recent progress in sign language translation utilizing these methods.

2.2.2.1 Text-to-Gloss

Text-to-gloss, an instantiation of sign language translation, is the task of translating between a spoken language text and sign language glosses. It is an appealing area of research because of its simplicity for integrating in existing NMT pipelines, despite recent works such as Yin and Read (2020) and Müller et al. (2023) claiming that glosses are an inefficient representation of sign language, and that glosses are not a complete representation of signs (Pizzuto et al., 2006).

Zhao et al. (2000) used a Tree Adjoining Grammar (TAG)-based system to translate English sentences to American Sign Language (ASL) gloss sequences. They parsed the English text and simultaneously assembled an ASL gloss tree, using Synchronous TAGs (S. M. Shieber and Schabes, 1990; S. Shieber, 1994), by associating the ASL elementary trees with the English elementary trees and associating the nodes at which subsequent substitutions or adjunctions can occur. Synchronous TAGs have been used for machine translation between spoken languages (Abeillé et al., 1991), but this was the first application to a signed language.

Othman and Jemni (2012) identified the need for a large parallel sign language gloss and spoken language text corpus. They developed a part-of-speech-based grammar to transform English sentences from the Gutenberg Project ebooks collection (Lebert, 2008) into American Sign Language gloss. Their final corpus contains over 100 million synthetic sentences and 800 million words and is the most extensive English-ASL gloss corpus we know of. Unfortunately, it is hard to attest to the quality of the corpus, as the authors did not evaluate their method on real English-ASL gloss pairs.

Egea Gómez et al. (2021) presented a syntax-aware transformer for this task, by injecting word dependency tags to augment the embeddings inputted to the encoder. This involves minor modifications in the neural architecture leading to negligible impact on computational complexity of the model. Testing their model on the RWTH-PHOENIX-Weather-2014T (Camgöz et al., 2018), they demonstrated that injecting this additional information results in better translation quality.

2.2.2.2 Gloss-to-Pose

Gloss-to-pose, subsumed under the task of sign language production, is the task of producing a sequence of poses that adequately represent a sequence of signs written as gloss.

To produce a sign language video, Stoll et al. (2018) construct a lookup table between glosses and sequences of 2D poses. They align all pose sequences at the neck joint of a reference skeleton and group all sequences belonging to the same gloss. Then, for each group, they apply dynamic time warping and average out all sequences in the group to construct the mean pose sequence. This approach suffers from not having an accurate set of poses aligned to the gloss and from unnatural motion transitions between glosses.

To alleviate the downsides of the previous work, Stoll et al. (2020) construct a lookup table of gloss to a group of sequences of poses rather than creating a mean pose sequence. They build a Motion Graph (Min and Chai, 2012), which is a Markov process used to generate new motion sequences that are representative of natural motion, and select the motion primitives (sequence of poses) per gloss with the highest transition probability. To smooth that sequence and reduce unnatural motion, they use a Savitzky–Golay motion transition smoothing filter (Savitzky and Golay, 1964).

2.2.2.3 Pose-to-Video

Pose-to-video, also known as motion transfer or skeletal animation in the field of robotics and animation, is the conversion of a sequence of poses to a video. This task is the final “rendering” of sign language in a visual modality.

Chan et al. (2019) demonstrated a semi-supervised approach where they took a set of videos, ran pose estimation with OpenPose (Cao et al., 2021), and learned an image-to-image translation (Isola et al., 2017) between the rendered skeleton and the original video. They demonstrated their approach on human dancing, where they could extract poses from a choreography and render any person as if *they* were dancing. They predicted two consecutive frames for temporally coherent video results and introduced a separate pipeline for a more realistic face synthesis, although still flawed.

Wang et al. (2018) suggested a similar method using DensePose (Güler et al., 2018) representations in addition to the OpenPose (Cao et al., 2021) ones. They formalized a different model, with various objectives to optimize for, such as background-foreground separation and temporal coherence by using the previous two timestamps in the input.

Using the method of Chan et al. (2019) on "Everybody Dance Now", Ventura et al. (2020) asked, "Can Everybody Sign Now?" and investigated if people could understand sign language from automatically generated videos. They conducted a study in which participants watched three types of videos: the original signing videos, videos showing only poses (skeletons), and reconstructed videos with realistic signing. The researchers evaluated the participants' understanding after watching each type of video. The results of the study revealed that participants preferred the reconstructed videos over the skeleton videos. However, the standard video synthesis methods used in the study were not effective enough for clear sign language translation. Participants had trouble understanding the reconstructed videos, suggesting that improvements are needed for better sign language translation in the future.

As a direct response, Saunders et al. (2020) showed that like in Chan et al. (2019), where an adversarial loss was added to specifically generate the face, adding a similar loss to the hand generation process yielded high-resolution, more photo-realistic continuous sign language videos. To further improve the hand image synthesis quality, they introduced a keypoint-based loss function to avoid issues caused by motion blur.

In a follow-up paper, Saunders et al. (2021) introduced the task of Sign Language Video Anonymisation (SLVA) as an automatic method to anonymize the visual appearance of a sign language video while retaining the original sign language content. Using a conditional variational autoencoder framework, they first extracted pose information from the source video to remove the original signer appearance, then generated a photo-realistic sign language video of a novel appearance from the pose sequence. The authors proposed a novel style loss that ensures style consistency in the anonymized sign language videos.

2.2.3 Method

In this section, we provide an overview of our text-to-gloss-to-pose-to-video pipeline, detailing the components and how they work together to convert input spoken language text into a sign language video. The pipeline consists of three main components: text-to-gloss translation, gloss-to-pose conversion, and pose-to-video animation. For text-to-gloss translation, we provide three different alternatives: a lemmatizer, a rule-based word reordering and dropping component, and a neural machine translation system. Figure 2.9 illustrates the entire pipeline and its components.

Funded by the Horizon 2020
Framework Programme of the European Union

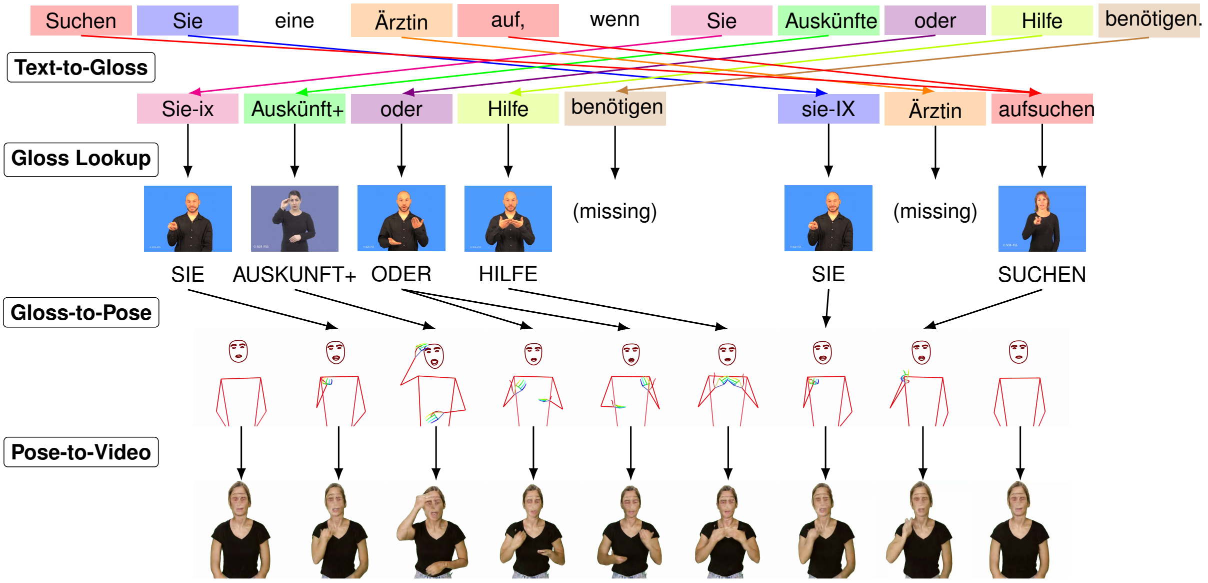


Figure 2.9: *Pipeline of the proposed text-to-gloss-to-pose-to-video approach for sign language translation. Starting with a German sentence, the system applies text-to-gloss translation, for example, using a rule-based word reordering and dropping component. The resulting gloss sequence is used to search for relevant videos from a lexicon of Swiss German Sign Language (DSGS). The poses of each relevant video are then extracted and concatenated in the gloss-to-pose step to create a pose sequence for the sentence, which is then transformed back to a (synthesized) video using the pose-to-video model. The figure demonstrates the transformation of the sentence "Suchen Sie eine Ärztin auf, wenn Sie Auskünfte oder Hilfe benötigen." ('Seek out a doctor if you need information or assistance.') to a sequence of glosses, the search for relevant videos for each gloss, the concatenation of pose videos, and the final video output.*

2.2.3.1 Pipeline

Below, we describe the high-level structure of our pipeline, including the text-to-gloss translation, gloss-to-pose conversion, and pose-to-video animation components:

1. **Text-to-Gloss Translation:** The input (spoken language) text is first processed by the text-to-gloss translation component, which converts it into a sequence of glosses.
2. **Gloss-to-Pose Conversion:** The sequence of glosses generated from the previous step is then used to search for relevant videos from a lexicon of signed languages (e.g., DSGS, LSF-CH, LIS-CH). We extract the skeletal poses from the relevant videos using a state-of-the-art pre-trained pose estimation framework. These poses are then cropped, concatenated, and smoothed, creating a pose representation for the input sentence.
3. **Pose-to-Video Generation:** The processed pose video is transformed back into a synthesized video using an image translation model, based on a custom training of Pix2Pix.

We note that while the pose-to-video generation is part of our pipeline, it will not be used in EASIER systems.

2.2.3.2 Implementation Details

Our system accepts spoken language text as input and outputs an *.mp4* video file, or a binary *.pose* file, which can be handled by the *pose-format* library (Moryossef and Müller, 2021) in Python and JavaScript. The *.pose* file represents the sign language pose sequence generated from the input text. To make our system easy to use, we deploy it as an HTTP endpoint that receives text as input and outputs the *.pose* file. We provide a demonstration of our system using `https://sign.mt`, with support for the three signed languages of Switzerland.

We implement our pipeline using Python and package it using Flask, a lightweight web framework. This allows us to create an HTTP endpoint for our application, making it easy to integrate with other systems and web applications. Our system is deployed on a Google Cloud Platform (GCP) server, providing scalability and easy access. Furthermore, we release the source code of our implementation as open-source software, allowing others to build upon our work and contribute to improving the accessibility of sign language translation systems.

By implementing our system as an open-source Python application and deploying it as an HTTP endpoint, we aim to facilitate collaboration and improvements to sign language translation systems.

2.2.4 Text-to-Gloss

We explore three different components as part of text-to-gloss translation, including a lemmatizer (Section 2.2.4.1), a rule-based word reordering and dropping component (Section 2.2.4.2), and a neural machine translation (NMT) system (Section 2.2.4.3).

2.2.4.1 Lemmatizer

We use the *Simplemma* simple multilingual lemmatizer for Python (Barbaresi, 2023). The lemmatizer reduces words to their base form (i.e., lemma), which is useful for our case, as it helps to preserve meaning while reducing the complexity of the input. This approach is limited by the use of the simplistic context-free lemmatizer, since no sense information is captured in the lemma, which causes ambiguity.

2.2.4.2 Word Reordering and Dropping

We generate near-glosses for sign language from spoken language text using a rule-based approach. The process from converting spoken language sentences into sign language gloss sequences can be naively summarized by a removal of word inflection, an omission of punctuation and specific words, and word reordering. To address these differences, we adopt the rule-based approach from Moryossef et al. (2021b) to generate near-glosses from spoken language: lemmatization of spoken words, PoS-dependent word deletion, and word order permutation. With their permission, we re-share these rules:

Specifically, we use spaCy (Montani et al., 2023) for lemmatization, PoS tagging and dependency parsing. Unlike Simplelemma, the spaCy lemmatizer is language specific and context based. We drop words that are not content words (e.g., articles, prepositions), as they are largely unused in signed languages, but keep possessive and personal pronouns as well as nouns, verbs, adjectives, adverbs, and numerals. We devise a short list of syntax transformation rules based on the grammar of the sign language and the corresponding spoken language. We identify the subject, verb, and object in the input text and reorder them to match the order used in the signed language. For example, for German-to-DGS, we reorder SVO sentences to SOV, move verb modifying adverbs and location words to the start of the sentence (a form of topicalization), move negation words to the end.

The specific rules we use for German to DGS/DSGS are:

1. For each subject-verb-object triplet $(s, v, o) \in \mathcal{S}$, swap the positions of v and o in $\mathcal{S}$
2. Keep all tokens $t \in \mathcal{S}$ if **PoS**$(t) \in$ {noun, verb, adjective, adverb, numeral, pronoun}
3. If **PoS**$(t) =$ adverb and **HEAD**$(t) =$ verb, move t to the start of S
4. If **NER**$(t) =$ location, move t to the start of S
5. If **DEP**$(t) =$ negation, move t to the end of S
6. Lemmatize all tokens $t \in \mathcal{S}$

We first split each sentence into separate clauses and reorder them before we apply these rules to each clause. Reordering the clauses may be needed for conditional sentences where the conditional subordinate clause should precede the main clause, as in “if…then…”. These rules allow us to transform spoken language text into near-glosses that more closely match the word order and structure of sign language. Overall, our rule-based approach provides a flexible and effective way to generate near-glosses for sign language from spoken language text, with the ability to incorporate language-specific rules to capture the nuances of different sign languages. This approach employs a more accurate lemmatizer, however, it still suffers from word sense ambiguity.

2.2.4.3 Neural Machine Translation

As an alternative to rule-based transformations of text to glosses, we consider a neural machine translation (NMT) system. We use an existing gloss translation system trained by EASIER and described in our earlier WP4 **Deliverable 4.2**.

The particular system we use was trained on the Public DGS Corpus. The model is multilingual, following the methodology described in Johnson et al. (2017) which inserts special tokens into all source sentences to indicate the desired target language. Therefore, the model can translate from German text to DGS glosses and vice versa. Our automatic evaluation in **Deliverable 4.2** confirmed that one multilingual system leads to higher translation quality than individual bilingual systems.

2.2.4.4 Language-Dependent Implementation

In this work, we study three sign languages: LIS-CH, LSF-CH and DSGS. For LIS-CH and LSF-CH we always apply our simple lemmatizer (Section 2.2.4.1) for the text-to-gloss step. The lemmatizer-only component is universally applicable to many more languages. However, it is worth noting that this approach does not capture the full spectrum of syntactic and morphological changes necessary in going from a spoken language to a sign language, which likely leads to suboptimal translations.

For DSGS, we explored different options for text-to-gloss, comparing the lemmatizer (Section 2.2.4.1), rule-based system (Section 2.2.4.2) and NMT system (Section 2.2.4.3). We observed that the glosses output by the NMT system are less accurate than rule-based reordering. A potential explanation for this is that the system is trained on German Sign Language (DGS) data. Due to the inherent differences between DGS and DSGS, using the NMT system could result in inaccurate translations or out-of-lexicon glosses. Furthermore, we found that the NMT system is not robust to out-of-domain text or capitalization differences, which further limits its applicability in these scenarios.

In the end, for DSGS we opted to employ our rule-based system (Section 2.2.4.2), which has been tailored to accommodate the unique linguistic characteristics of DSGS, and produces the best results.

2.2.5 Gloss-to-Pose

Gloss-to-pose translation involves converting sign language glosses into a sequence of poses that adequately represent a sequence of signs.

We use the SignSuisse dataset (Schweizerischer Gehörlosenbund SGB-FSS, 2023), which consists of sign language videos in three different languages. We extract skeletal poses from these videos using Mediapipe Holistic (Grishchenko and Bazarevsky, 2020), a state-of-the-art pose estimation framework that estimates 3D coordinates of various landmarks on the human body, including the face, hands, and body. We preprocess the poses by ensuring that the `body` wrists are in the same location as the `hand` wrists, removing the legs, hands, and face from the body pose, and cropping the videos in the beginning and end to avoid returning to a neutral body position.

We concatenate the poses for each gloss by finding the best 'stitching' point that minimizes L2 distance. We then concatenate these poses, adding 0.2 seconds of 'padding' in between, before applying cubic smoothing on each joint to ensure smooth transitions between signs, and filling in missing keypoints. Finally, we apply a Savitzky-Golay motion transition smoothing filter (Savitzky and Golay, 1964), similar to Stoll et al. (2020), to reduce unnatural motion.

2.2.6 Pose-to-Video

We use a semi-realistic human-like avatar system to animate the poses generated by our approach. The avatar system is a Pix2Pix model (Isola et al., 2017) adjusted to operate on pose sequences, not individual images. With her permission, we use the likeness of Maayan Gazuli[4]. We use OpenCV (Bradski, 2000) to render the poses as images and feed them into the Pix2Pix model to generate realistic-looking video frames. The avatar system can run in real-time on supported devices and is integrated into `https://sign.mt` (Moryossef, 2023). This system is far from the state of the art, however, we believe that the open-source nature of it will bring rapid improvements, like faster inference speed, and higher animation quality.

2.2.7 Conclusions

We presented an implementation of a text-to-gloss-to-pose-to-video pipeline for sign language translation, focusing on Swiss German Sign Language, Swiss French Sign Language, and Swiss Italian Sign Language. Our approach comprises three main components: text-to-gloss translation, gloss-to-pose conversion, and pose-to-video animation.

We explained the structure of our system and discussed its limitations, as well as future work directions to address them. These directions have the potential to improve our system, and we look forward to exploring them in collaboration with the open-source community.

The main contribution of this work is the creation of a reproducible baseline for spoken to signed language translation. The system should serve as a baseline for comparison with more sophisticated sign language translation systems and can be improved upon by the community. Our systems for the three signed languages of Switzerland are available on `https://sign.mt`.

2.2.8 Future Work

Here we include several future work directions that we believe have the potential to further enhance the performance and user experience of our system for text-to-gloss-to-pose-to-video

[4] https://nlp.biu.ac.il/~amit/datasets/GreenScreen/

generation, and we look forward to exploring these possibilities in the future, together with the open-source community.

Qualitative Evaluation To evaluate the effectiveness of our approach, we will conduct a study to gather first impressions from deaf users. We already recruited a group of deaf individuals and will ask them to use our system to translate text into sign language videos.

Each participant will be asked to provide feedback on the system after using it to translate five different sentences from German into DSGS. We will provide the sentences to the participants, and they will be asked to sign the translations generated by our system. After each sentence, the participant will be asked to provide feedback on the accuracy of the translation, the quality of the poses and/or synthesized video, and the overall usability of the system.

Gloss Sense Disambiguation The current approach to text-to-gloss translation relies on a simple lemmatizer and a rule-based word reordering and dropping component, which can lead to ambiguity in the glosses produced. In the future, we can enhance our system by incorporating gloss sense disambiguation to better capture the intended meaning of the input text. Our NMT approach responds with gloss IDs from the MeineDGS corpus, which already are sense-disambiguated. Annotation of our sign language lexicon with senses will allow us to retrieve the relevant sense.

Handling Unknown Glosses Where we encounter a gloss that does not exist in our lexicon, we propose exploring alternative methods to generate a video for it. One possible solution is to leverage another lexicon that includes a written representation of the gloss in question (e.g., SignWriting (Sutton, 1990) or HamNoSys (Prillwitz and Zienert, 1990)), or to employ a neural machine translation system to translate the individual concept to a writing system. Utilizing the capabilities of machine translation to embed words, we can perform a fuzzy match, addressing issues such as synonyms.

Additionally, for named entities such as proper nouns and place names that are not covered by our current gloss-to-pose conversion system, we could revert to fingerspelling them.

Once we have the written representation, we can use a system like Ham2Pose (Shalev-Arkushin et al., 2023) to generate a single sign video from the writing. When combined with fingerspelling for named entities, this approach should enable greater coverage of the language.

Handling Unknown Gloss Variations In situations where the required gloss variation is not present in the lexicon but a related gloss exists, we propose developing a system that can modify the known gloss to generate the desired variation. This would allow for better handling of unknown gloss variations and increase the accuracy of the information conveyed by the signing.

Number Forms For words like *KINDER* (children), we may encounter glosses such as *KIND+*, which represent “child" in plural form. Assuming that we have *KIND* in our lexicon but not *KINDER*, a system could be developed to modify signs to plural forms, such as by repeating

Funded by the Horizon 2020
Framework Programme of the European Union

movements or incorporating specific handshapes or locations that indicate plurality in the target sign language. Conversely, if we only have the plural form of a gloss in our lexicon, the system could be designed to generate the singular form by removing or modifying the elements that indicate plurality.

Part-of-Speech Conversion Another challenge arises when nouns or verbs exist in the lexicon, but their counterparts do not. For instance, if *HELFEN* (to help) is present in the dictionary as a verb, but *HILFE* (help) does not exist as a noun, a system could be designed to modify signs from one part of speech to another, such as from verb to noun or noun to verb. This system could potentially involve morphological or movement modifications, depending on the linguistic rules of the target sign language.

Post-editing Pose Sequences The current approach generates a sequence of poses that represent a sign language sentence. We believe that there is also room for improvement in terms of the fluency and naturalness of the generated sequence. Exploring the use of automatic post-editing techniques is necessary. One such approach could identify datasets that include sentences and gloss sequences, such as the Public DGS Corpus, then, using our gloss-to-pose approach generate a pose sequence with poses from the lexicon, and could learn a diffusion model between the synthetic and real pose sequences.

2.2.9 Interface with Avatar in WP2

For the spoken-to-signed translation systems we developed, we envision that their output is fed to an avatar system such as the one developed by WP2. There are several ways in which the translation output can be visualized, including interfacing with an avatar.

- For all five language pairs, we offer a text-to-gloss component. Depending on the language pair, the component is either a full-fledged translation system (for DE→DGS and EN→BSL, described in **Deliverable 4.2**) or a simpler transformation based on lemmatization or rules (for DE→DSGS, FR→LSF and IT→LIS, described in Section 2.2.4). The glosses of this component can be represented by the EASIER avatar as developed in WP2, after a look-up process in a Gloss-HamNoSys dictionary and incorporation of prosodic information relevant to the translated text along with emotion features deriving from WP7.
- For three language pairs we deliver a more comprehensive text-to-pose translation system (for DE→DSGS, FR→LSF and IT→LIS, described in Sections 2.2.4 and 2.2.5).

Therefore, translation output visualization may be achieved either via an avatar system which could take glosses as an input and provide 3D signed representations of the translations, or via pose sequences displayed on 2D stick figures.

Funded by the Horizon 2020
Framework Programme of the European Union

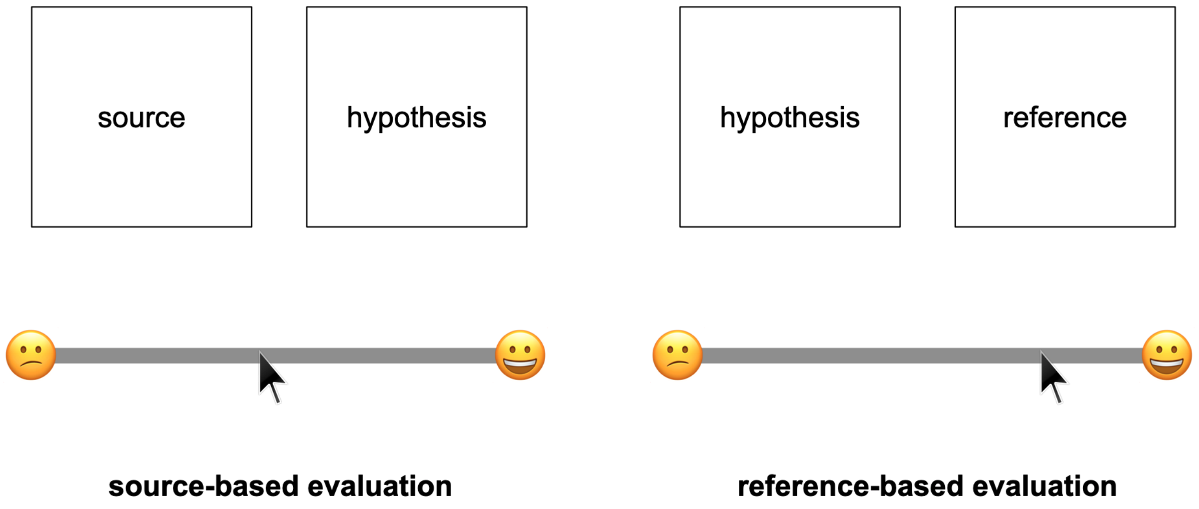


Figure 2.10: *Simplified illustration of **direct assessment** methods, widely used protocols for human evaluations of machine translation systems. source=input to the translation system. hypothesis=output of the system. reference=human translation.*

2.2.10 Interim evaluation

We conduct a preliminary evaluation of translation quality for some of the spoken-to-signed translation systems described in this section. Specifically, we evaluate the language pairs DE→DSGS and IT→LIS. For DE→DSGS, we evaluate two different text-to-gloss components that are based on simple lemmatization or hand-written rules, respectively (see Section 2.2.4). For IT→LIS we only evaluate simple lemmatization.

Section 2.2.10.1 outlines evaluation designs that are common in machine translation research, and the exact design chosen for this interim evaluation. Section 2.2.10.2 summarizes the outcome of the evaluation.

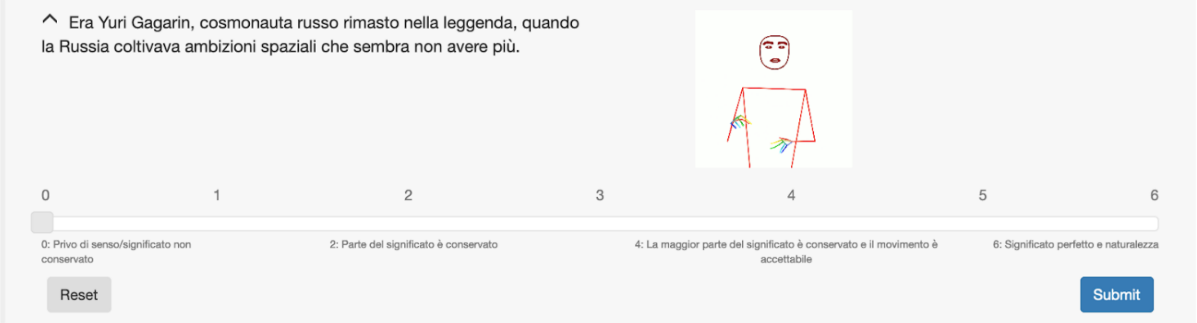
2.2.10.1 Human evaluation protocol

Below we outline considerations for a human evaluation of sign language machine translation, as there is hardly any previous study to build upon.

Common evaluation protocols Human evaluations of machine translation output always have a *comparative* methodology, but individual methods vary in what is shown to an evaluator at any given time. The two most widely used methods are:

- **Direct assessment (DA):** One system is evaluated at any given time. The evaluator is asked to compare the MT output to either 1) the source or 2) the human reference translation. These sub-types are called *source-based* and *reference-based* DA, respectively.
- **Ranking:** several systems are evaluated at any given time. The evaluator is asked to sort system outputs by quality, producing a system ranking.

Funded by the Horizon 2020 Framework Programme of the European Union



Era Yuri Gagarin, cosmonauta russo rimasto nella leggenda, quando la Russia coltivava ambizioni spaziali che sembra non avere più.

0
1
2
3
4
5
6
0: Privo di senso/significato non conservato
2: Parte del significato è conservato
4: La maggior parte del significato è conservato e il movimento è accettabile
6: Significato perfetto e naturalezza
Reset
Submit

Figure 2.11: *Screenshot of Appraise, a browser-based tool for human evaluation of machine translation systems. The figure shows the specific view shown to evaluators for the interim evaluation.*

In recent years most MT evaluations have exclusively used DA methodology (Graham et al., 2016). Evaluators are shown either the source or the reference translation, and are asked to rate translation quality on a scale of 1 to 100. See Figure 2.10 for an illustration.

Design for EASIER interim evaluation Translation quality was assessed with source-based DA (the most ideal form of MT evaluation). We conducted an online study using the tool *Appraise* (Federmann, 2018). As suitable user interfaces are important for such evaluations (Grundkiewicz et al., 2021), the tool was adapted to sign language in many respects.

The tool was extended to support videos as an additional modality of translation inputs or outputs and to support evaluator instructions in a sign language. See Figure 2.11 for an example of the evaluator view of Appraise. This new version of Appraise was developed for the WMT-SLT shared task on sign language translation carried out by members of EASIER[5].

Requirements for human experts Ideally, evaluators for source-based DA are bilingual, and most proficient in the target language that the MT system produces. In the context of a sign language evaluation, this often means that individuals are Deaf sign language users for spoken-to-sign systems (assuming that the signers themselves state having higher proficiency in a sign language) and hearing sign language users with a spoken language as a first language for sign-to-spoken systems.

Instructions for evaluators The instructions for evaluators are adapted specifically to sign language and the new modalities (other than text) involved.

The instructions provide some guidance in the form of discrete quality levels (referred to as Scalar Quality Metric (SQM) (Freitag et al., 2021)) that partition the continuous scale of 1 to 100. The quality levels range from *"0 - No Meaning Preserved"* to *"6 - Perfect Meaning"*. See Figure 2.11 for an example how the quality levels are displayed to the user.

For spoken-to-sign evaluations where the output is a sign language video (or similar), we added an evaluation criterion specific to sign languages: *naturalness of motion*. We aim to distinguish

[5] https://www.wmt-slt.com/

between robotic and human-like, natural motion in system outputs.

Also, following the recent evaluations at the workshop for spoken language machine translation (IWSLT 2022; Anastasopoulos et al., 2022), we remove any mention of "grammar" from the descriptions of quality levels. This was done to shift attention away from grammatical issues in the target language towards translation-breaking differences in meaning. And similar to the domain of speech, our evaluation material features continuous signing, rather than formalized signing equivalent to a written text.

The full instructions for spoken-to-sign and sign-to-spoken evaluations are included in Appendix 5. We also translated these instructions to other spoken and signed languages, since Appraise also supports video instructions.

Extent of the evaluation We evaluate the language pairs DE→DSGS and IT→LIS. For DE→DSGS, we evaluate two different text-to-gloss components that are based on simple lemmatization or hand-written rules, respectively (see Section 2.2.4). For IT→LIS we only evaluate simple lemmatization. For each included system we evaluate 100 sentences, presenting them to evaluators in the original order (i.e. taking into account document context). For each language pair, two evaluators completed our evaluation.

Since the systems output pose sequences, we applied pose estimation to the human references and also display them as a pose sequence. This is to de-emphasize the difference between the poses and a real person doing the signing.

2.2.10.2 Outcome of interim evaluation

The results are shown in Tables 2.7 and 2.8. Transforming human reference translations to pose sequences noticeably reduces the quality scores assigned by evaluators for LIS but not for DSGS. This indicates that evaluators have different views about the legibility and acceptability of poses.

Moreover, for DSGS, simple lemmatization appears to perform better than our hand-written, linguistic rules. One potential explanation is that the hand-written rules, while being precise, are lacking coverage. However, the outcome should be taken with a grain of salt, as only 100 sentences were evaluated overall, by only two evaluators.

Overall, the general level of translation quality of our spoken-to-signed translation systems is much higher than for signed-to-spoken systems evaluated previously. For instance, Müller et al. (2022) conclude that the average translation quality of the best-performing signed-to-spoken system is 2 out of 100, while the spoken-to-signed systems evaluated here achieved a score between 10 out of 100 and 20 out of 100.

Funded by the Horizon 2020
Framework Programme of the European Union

Rank	Average score	System
1	25.914	HUMAN
2	10.223	simplemma

Table 2.7: *Average score given by human evaluators in our interim evaluation for IT→LIS. All systems (including the human translation) are in different quality clusters, as determined by a Mann-Whitney-U significance test.*

Rank	Average score	System
1	88.759	HUMAN
2	19.657	simplemma
3	14.255	rules

Table 2.8: *Average score given by human evaluators in our interim evaluation for DE→DSGS. All systems (including the human translation) are in different quality clusters, as determined by a Mann-Whitney-U significance test.*

Funded by the Horizon 2020
Framework Programme of the European Union

3 SPOKEN-TO-SPOKEN TRANSLATION (TASK 4.3)

Task 4.3 provides the model to translate between all six of the spoken languages in the EASIER project (English, French, Italian, German, Dutch and Greek). Furthermore, we conduct experiments to alleviate a form of gender bias in our models with a) data labelled with the gender of the speaker and b) controlled generation to improve the translation of gender specific terms relating to the speaker (English-Italian only).[6,7]

Parts of the research in this section was published in Lu et al. (2023). The text was adapted to fit into the context of this deliverable.

3.1 GENDER BIAS

Gender bias in machine translation comprises a number of different issues, the most prominent are:

1. stereotyping: when certain activities, occupations or professions are associated with gender, e.g. when a model defaults to use a male forms for *doctor* but female forms for *nurse* (Stanovsky et al., 2019)
2. speaker gender: when translating into languages that mark the gender of the speaker in certain contexts, models tend to default to the more commonly seen gendered forms (Vanmassenhove et al., 2018)
3. pronoun translation: models tend to have a bias towards the more frequently seen pronouns in training. Additionally, stereotyping, as listed above under 1), can affect pronoun translation (Loáiciga et al., 2017; Jwalapuram et al., 2020)

For this deliverable, we focus on the second kind of gender bias in the list, where translating into a language that marks the gender of the speaker openly can lead to gender ambiguity and potentially wrong translations (e.g. *I am happy* → *Je suis hereux/hereuse.*).

3.2 DATA

3.2.1 Europarl with Speaker Information

This dataset consists of European Parliament discussions (Koehn, 2005) annotated with meta information (Vanmassenhove and Hardmeier, 2018), including the gender of the speaker. This corpus contains a substantial number of first-person sentences, which makes this type of data particularly well suited to test gender bias in reference to the speaker.

[6] Code for fine-tuning the multilingual models: https://github.com/a-rios/ats-models

[7] Code for the bilingual English-Italian experiments: https://github.com/tianshuailu/debias_FUDGE

Funded by the Horizon 2020
Framework Programme of the European Union

	Europarl	
	sentences	M:F ratio
Italian	1.30M	2.07:1
French	1.44M	2.05:1
Greek	0.92M	2.03:1
German	1.30M	2.05:1
Dutch	1.42M	2.06:1

	ParlaMint (Italian)	
	sentences	M:F ratio
total	996.5k	2.5:1
filtered	91.6k	1:1

Table 3.1: *Overview on datasets. Europarl data is parallel with English for all languages.*

3.2.2 Italian ParlaMint

We further experiment with controlled generation on the English-Italian direction. For this set of experiments, in addition to the English-Italian Europarl data, we need monolingual Italian data to train classifiers. In order to keep the domain consistent, we use the Italian part of ParlaMint2.1 (Erjavec et al., 2021). We split the full speech segments into sentences to keep the units consistent with Europarl.

In Italian, adjectives and participles are marked with the gender of the speaker in certain grammatical contexts. The full dataset is quite large, and the utterances where the gender of the speaker is openly marked are relatively sparse. We therefore filter out sentences that do not contain neither adjectives nor participles, since these cannot be marked for the gender of the speaker, and thus provide no information to the classifiers. The size of the original and the filtered data set are shown in Table 3.1.

The ratio of male to female speakers throughout all data sets is roughly 2:1, however, with the filtered ParlaMint data, we use the same amount of utterances for both genders to ensure balanced positive and negative class sizes when training classifiers (Lu et al., 2023).

3.3 MODELS

3.3.1 Multilingual Machine Translation with Gender Tags

We fine-tune a multilingual version of LongT5, mLongT5 (Uthus et al., 2023), on the parallel Europarl data for our experiments. We use mLongT5 instead of mt5 (Xue et al., 2021) since it trains considerably faster.[8]

To save memory, we reduce the full vocabulary of >250k items to the 30k most frequent items in the languages in our data: We use the tokenizer to split all the training data plus 2M sentences from crawled news corpora and keep the most frequent 30k items as the model's vocabulary.[9]

[8] We use the *base* configuration of the pretrained model with 12 layers in encoder and decoder.

[9] The corpora used to create the vocabulary lists are:

- English, German, French, Italian: news crawl data from WMT'22: `https://data.statmt.org/news-crawl/`
- Greek, Dutch: CCaligned corpus data (El-Kishky et al., 2020): `https://opus.nlpl.eu/CCAligned.php`

	Baseline	with Gender Tags
en-el	20.3	**21.6**
el-en	31.5	**32.9**
en-it	26.5	**27.0**
it-en	31.7	**32.1**
en-fr	32.1	**32.6**
fr-en	35.4	**35.6**
en-nl	24.1	**24.5**
nl-en	29.8	**30.2**
en-de	22.9	**23.4**
de-en	31.1	**31.4**

Table 3.2: *BLEU scores of multilingual translation with and without gender annotations (beam=4).*

We train a standard multilingual model as baseline where the translation direction is controlled through an instruction on the source side.[10]

We fine-tune a second model with additional gender tags, following work by Vanmassenhove et al. (2018). Data splits, preprocessing and hyperparameters are identical to the baseline model, the only difference is a tag ($< masc >$ or $< fem >$) prefixed to the source text.

Both models are English-centric, i.e. the translation directions seen in training are English-to and from-English for each of the languages. The models can translate in the unseen directions as well (zero-shot), but the quality is below the translations from and to English. Unfortunately, multi-way parallel corpora annotated for speaker gender are not available at this time and we can therefore not test our models in the zero-shot directions.

3.3.2 Results

Table 3.2 shows that adding gender tags slightly, but consistently, improves BLEU scores across all translation directions.[11] This finding is in accordance with previous publications on this topic (Vanmassenhove et al., 2018).

In the following sections, we dive deeper into speaker gender issues for one of the translation directions (English-Italian).

3.4 ENGLISH→ITALIAN TRANSLATION WITH CONTROLLED GENERATION

We conduct more in-depth experiments on the translation direction English→Italian, this work was published in Lu et al. (2023).

[10] "Translate to X:"

[11] sacrebleu: “nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.1“

3.4.1 FUDGE

Yang and Klein (2021) propose Future Discriminators for Generation (FUDGE), a flexible and modular way of conditioning a generative model $\mathcal{G}$ on a desired attribute a that only requires access to the output probabilities of trained generation model $\mathcal{G}$.

FUDGE achieves this by training a binary classifier that predicts at each time step t whether the attribute a will be satisfied in the complete sequence, based on the already generated tokens $y_0 - y_t$.

A standard auto-regressive model predicts tokens based the previous steps:

$$P(X) = \prod_{i=1}^{n} P(x_i|x_{1:i-1}) \tag{3.1}$$

With FUDGE, the prediction is based on an additional feature a:

$$P(X) = \prod_{i=1}^{n} P(x_i|x_{1:i-1}, a) \tag{3.2}$$

Which can be formulated as:

$$P(x_i|x_{1:i-1}, a) \propto P(a|x_{1:i-1})P(x_i|x_{1:i-1}) \tag{3.3}$$

The first term in this equation is modelled by a classifier, whereas for the second term, any auto-regressive generation model can be used. Finally, the weight of the classifier's contribution to the prediction is controlled through a hyperparameter λ.

3.4.2 Generation Models $\mathcal{G}$ and $\mathcal{G}_t$

For our experiments, we train two English→Italian translation models, $\mathcal{G}$ and $\mathcal{G}_t$. We train both models on the same sentence pairs, with the exception that $\mathcal{G}_t$'s data set includes gender tags on the English source side, similar to the multilingual experiments described in section 3.3.

We fine-tune mt5 (Xue et al., 2021) on the English→Italian part of the gender annotated Europarl corpus (Vanmassenhove and Hardmeier, 2018) introduced in section 3.2.1. Similar to the multilingual experiments in the previous section, we reduce the vocabulary to the 25k most frequent entries in Italian and English. $\mathcal{G}$ is mt5 fine-tuned directly on the Europarl data, whereas $\mathcal{G}_t$ is mt5 fine-tuned on Europarl data annotated with gender tags. Data splits and preprocessing are identical for $\mathcal{G}$ and $\mathcal{G}_t$.

3.4.3 Classifiers $\mathcal{B}_f$ and $\mathcal{B}_m$

The attributes we want FUDGE to predict are feminine and masculine gender of the speaker. We train two classifiers $\mathcal{B}_f$ and $\mathcal{B}_m$ on the monolingual ParlaMint data described in section 3.2.2. Each of these classifiers is then combined with the translation models $\mathcal{G}$ and $\mathcal{G}_t$, resulting in four combinations, as illustrated in Figure 3.1.

An advantage of FUDGE is the fact that it only needs access to the output logits of the generator model, meaning $\mathcal{G}$ and $\mathcal{G}_t$ can be directly combined with $\mathcal{B}_f$ and $\mathcal{B}_m$ without additional fine-tuning or modification. This allows us to directly use $\mathcal{G}$ and $\mathcal{G}_t$ as baselines (Lu et al., 2023).

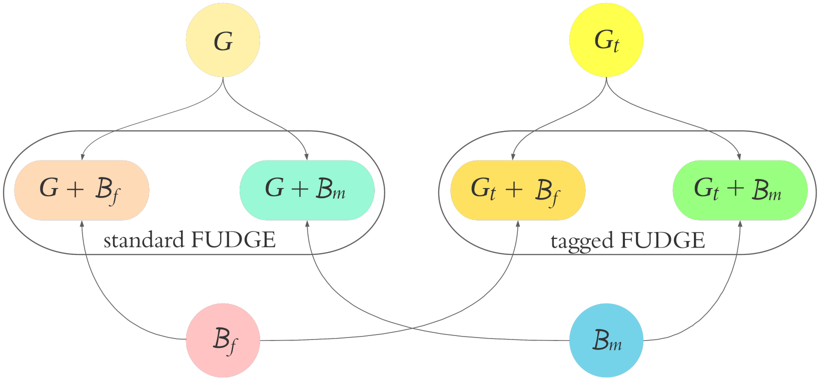


Figure 3.1: *Illustration of four combinations between the underlying translation models $\mathcal{G}$ (translation model trained on original data sets), $\mathcal{G}_t$ (translation model trained on tagged data sets) and two classifiers $\mathcal{B}_f$ (feminine), $\mathcal{B}_m$ (masculine) (Lu et al., 2023).*

3.4.4 Evaluation

In addition to the standard metric BLEU (Papineni et al., 2002), we evaluate on MuST-SHE v1.2, a multilingual challenge set that allows for a fine-grained analysis of gender bias in Machine and Speech Translation (Savoldi et al., 2022). MuST-SHE v1.2 contains 656 first-person sentences out of 1073 that we use in our evaluation.

For word-level evaluation, MuST-SHE performs a fine-grained qualitative analysis of the system's accuracy in producing the target gender-marked words. MuST-SHE computes the accuracy as the proportion of gender-marked words in the references that are correctly translated by the system. An upper bound of one match for each gender-marked word is applied to prevent rewarding over-generated terms.

For agreement-level evaluation, MuST-SHE inspects the agreement chain coverage and translation accuracy. Each agreement chain is composed of several agreement terms. The agreement chain is in coverage only when all the terms appear in the translation (regardless of their gender forms). Then MuST-SHE further evaluates the accuracy of the in-coverage chains. Either the agreement is not respected, i.e. inconsistent (*Inc*), or it is respected with the correct gender (*Correct*) or wrong gender (*Wrong*).

Funded by the Horizon 2020
Framework Programme of the European Union

	StandardFUDGE ($\mathcal{G}$ + $\mathcal{B}_f$)			TaggedFUDGE ($\mathcal{G}_t$ + $\mathcal{B}_f$)		
	Verbs	Nouns	Adj-des	Verbs	Nouns	Adj-des
$baseline$	27.4	11.4	35.4	27.3	13.5	36.3
$\lambda = 1$	43.7	12.8	42.9	39.5	13.2	45.7
$\lambda = 2$	60.6	13.2	61.2	56.3	**20.5**	55.1
$\lambda = 3$	62.1	10.8	55.1	63.6	14.3	61.7
$\lambda = 4$	70.1	11.8	61.2	**67.1**	15.4	64.6
$\lambda = 5$	**71.0**	**17.1**	**61.4**	62.9	19.0	**66.0**

Table 3.3: *Open-class POS accuracy on feminine forms of FUDGE $\mathcal{B}_f$, combined with standard baseline $\mathcal{G}$ and tagged baseline $\mathcal{G}_t$.*

	StandardFUDGE ($\mathcal{G}$ + $\mathcal{B}_m$)			TaggedFUDGE ($\mathcal{G}_t$ + $\mathcal{B}_m$)		
	Verbs	Nouns	Adj-des	Verbs	Nouns	Adj-des
$baseline$	87.8	**97.6**	94.3	94.4	**97.6**	94.1
$\lambda = 1$	91.4	96.3	**94.4**	94.5	97.5	92.2
$\lambda = 2$	92.9	97.5	94.2	95.8	97.5	91.7
$\lambda = 3$	94.1	97.4	94.1	93.1	97.5	92.2
$\lambda = 4$	**96.9**	97.5	94.1	**97.0**	97.3	**96.1**
$\lambda = 5$	96.6	97.5	92.0	95.5	97.5	91.8

Table 3.4: *Open-class POS accuracy on masculine forms of FUDGE $\mathcal{B}_m$, combined with standard baseline $\mathcal{G}$ and tagged baseline $\mathcal{G}_t$.*

We use a beam size of 4 in all our experiments.

3.4.5 Results

The hyperparameter λ determines how much weight is accorded to the classifier over the generation model's predictions during inference. We test each model with λ ranging from 1 to 5.

Tables 3.3 and 3.4 display the word-level open-class POS accuracy of standard FUDGE and tagged FUDGE with λ ranging from 1 to 5. *Adj-des* denotes descriptive adjectives.[12] As shown in Table 3.3, for both standard and tagged FUDGE, the accuracy of all three feminine form open-class words improves with the increase of λ. On the masculine forms on the other hand, the baselines are already very good and improvements with FUDGE are negligible, see Table 3.4.

Tables 3.5 and 3.6 illustrate the feminine and masculine gender agreement evaluation results of standard FUDGE and tagged FUDGE with λ ranging from 1 to 5. Table 3.5 shows that the tagged baseline ($\mathcal{G}_t$) generates the least inconsistent agreement chains, but combining this model with FUDGE $\mathcal{B}_f$ can still improve the number of chains with the correct gender. On the other hand, both baselines are very good at generating correct and consistent agreement chains for masculine speakers, as shown in Table 3.6, improvements with FUDGE are relatively

[12] E.g. *Sono **certo/a** che..* - I'm **certain** that...

Funded by the Horizon 2020
Framework Programme of the European Union

	StandardFUDGE ($\mathcal{G}$ + $\mathcal{B}_f$)			TaggedFUDGE ($\mathcal{G}_t$ + $\mathcal{B}_f$)		
	$Correct \uparrow$	$Wrong \downarrow$	$No \downarrow$	$Correct \uparrow$	$Wrong \downarrow$	$Inc \downarrow$
$baseline$	45.5	36.4	18.2	48.6	37.1	**14.3**
$\lambda = 1$	52.8	33.3	13.9	45.7	34.3	20.0
$\lambda = 2$	57.9	28.9	**13.2**	52.6	31.6	15.8
$\lambda = 3$	52.8	27.8	19.4	**56.7**	**27.0**	16.2
$\lambda = 4$	57.1	20.0	22.9	51.3	32.4	16.2
$\lambda = 5$	**63.6**	**18.2**	18.2	44.7	34.2	21.1

Table 3.5: *Accuracy on feminine gender agreement chains of FUDGE $\mathcal{B}_f$, combined with standard baseline $\mathcal{G}$ and tagged baseline $\mathcal{G}_t$. Wrong=agreement consistent in chain, but wrong gender, Inc = inconsistent gender in chain.*

	StandardFUDGE ($\mathcal{G}$ + $\mathcal{B}_m$)			TaggedFUDGE ($\mathcal{G}_t$ + $\mathcal{B}_m$)		
	$Correct \uparrow$	$Wrong \downarrow$	$No \downarrow$	$Correct \uparrow$	$Wrong \downarrow$	$Inc \downarrow$
$baseline$	91.1	3.6	5.4	**96.2**	**0.0**	3.8
$\lambda = 1$	94.5	1.8	3.6	94.4	1.9	3.7
$\lambda = 2$	94.4	1.9	3.7	94.2	1.9	3.8
$\lambda = 3$	94.4	1.9	3.7	94.4	1.9	3.7
$\lambda = 4$	**96.5**	**0.0**	**3.5**	**96.2**	**0.0**	3.7
$\lambda = 5$	92.3	**0.0**	7.7	94.7	1.8	**3.5**

Table 3.6: *Accuracy on masculine gender agreement chains of FUDGE $\mathcal{B}_m$, combined with standard baseline $\mathcal{G}$ and tagged baseline $\mathcal{G}_t$. Wrong=agreement consistent in chain, but wrong gender, Inc = inconsistent gender in chain.*

small.

Table 3.7 shows the BLEU scores of standard FUDGE and tagged FUDGE with both feminine and masculine classifiers, i.e. the four models illustrated in Figure 3.1. BLEU scores of both standard and tagged FUDGE decreases with the increase of λ. Since the classifiers were trained on a relatively small amount of data compared to the generation models, their fluency and grammaticality is not as good. There is a trade-off between according the classifiers more weight to correct gender mistakes, but also maintain the fluency and grammaticality of the mT5 baselines. With higher values for λ, the classifier starts to over-correct predictions, for an example see Table 3.8.

	StandardFUDGE ($\mathcal{G}$ + $\mathcal{B}_m$)		TaggedFUDGE ($\mathcal{G}_t$ + $\mathcal{B}_m$)	
	feminine	masculine	feminine	masculine
baseline	**27.2**	**27.0**	**27.5**	**27.1**
$\lambda = 1$	27.1	**27.0**	27.3	26.9
$\lambda = 2$	27.0	26.8	27.2	26.9
$\lambda = 3$	26.9	26.7	27.0	26.7
$\lambda = 4$	26.5	26.6	26.6	26.5
$\lambda = 5$	26.2	26.4	26.2	26.5

Table 3.7: *BLEU scores of standard FUDGE and tagged FUDGE with both feminine and masculine classifiers.*

1)	female speaker:
en	*I am* ***sure*** *you will agree ...*
reference	*Sono* ***certa*** *che sarà d'accordo ...*
baseline	*Sono* ***sicuro*** *che lei concorderà ...*
FUDGE	*Sono* ***sicura*** *che lei concorderà ...*
2)	female speaker:
en	*The internet is a* ***medium*** *...*
reference	*Internet è un* ***mezzo*** *...*
baseline	*Internet è un* ***medio*** *...*
FUDGE	*Internet è un* ***media*** *...*

Table 3.8: *Fudge examples: 1) desired correction 2) over-correction resulting in a wrong translation.*

4 CONCLUSIONS AND FUTURE WORK

This deliverable summarizes the progress made in the EASIER project towards the general goal of improving sign language machine translation. Much of our research was dedicated to empirical comparisons between different approaches, some known from literature, some entirely novel, that have never been compared on a large scale. The best systems are delivered for use in the EASIER project and for the general public.

For instance, we test a series of signed-to-spoken translation systems, varying the sign language representation used to encode the source language. Representations we evaluated include pose estimation systems, and an array of learned continuous (numerical) or discrete representations called EMSL. For spoken-to-signed translation, we dedicated much of our attention to comparing different approaches for text-to-pose translation. We also released our implementations, which are the first open-source implementations that are available and that could be used as baselines for others.

Furthermore, we deliver models that can translate between all the spoken languages in the EASIER project (English, Greek, Dutch, Italian, German, French). For one of the language pairs, English→Italian, we conduct a more in-depth set of experiments on reducing gender bias with controlled generation, which improves translation accuracy of the rarer feminine forms considerably.

Future work in the EASIER project During the remainder of the project, WP4 will be concerned with

- **More spoken-to-signed systems:** We will develop additional systems for the language pairs DE→DGS and EN→BSL, using the methods described in Section 2.2.
- **Final evaluation:** In the following -months a large-scale human evaluation of the EASIER translation systems will be conducted. In this evaluation we will show to professional translators the systems that performed best according to automatic quality metrics. The evaluation covers all five language pairs, in both translation directions.
- **Quality estimation:** In the final months of the project WP4 will develop and contribute a quality estimation system capable of predicting how good an automatic translation is.

More generally, beyond the end of the EASIER project, we believe more research is needed on creating better benchmark datasets, on developing basic NLP tools such as segmentation, and automatic metrics for spoken-to-signed translation.

REFERENCES

Abeillé, Anne, Yves Schabes, and Aravind K Joshi (1991). *Using lexicalized tags for machine translation*. Tech. rep. MS-CIS-91-44. University of Pennsylvania Department of Computer and Information Sciences.

Albanie, Samuel, Gül Varol, Liliane Momeni, Hannah Bull, Triantafyllos Afouras, Himel Chowdhury, Neil Fox, Bencie Woll, Rob Cooper, Andrew McParland, and Andrew Zisserman (2021). "BOBSL: BBC-Oxford British Sign Language Dataset". In: *CoRR* abs/2111.03635. arXiv: 2111.03635.

Anastasopoulos, Antonios, Loic Barrault, Luisa Bentivogli, Marcely Zanon Boito, Ondřej Bojar, Roldano Cattoni, Anna Currey, Georgiana Dinu, Kevin Duh, Maha Elbayad, Clara Emmanuel, Yannick Estève, Marcello Federico, Christian Federmann, Souhir Gahbiche, Hongyu Gong, Roman Grundkiewicz, Barry Haddow, Benjamin Hsu, Dávid Javorský, Věra Kloudová, Surafel Lakew, Xutai Ma, Prashant Mathur, Paul McNamee, Kenton Murray, Maria Nădejde, Satoshi Nakamura, Matteo Negri, Jan Niehues, Xing Niu, John Ortega, Juan Pino, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Yogesh Virkar, Alexander Waibel, Changhan Wang, and Shinji Watanabe (2022). "Findings of the IWSLT 2022 Evaluation Campaign". In: *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*. Dublin, Ireland (in-person and online): Association for Computational Linguistics, pp. 98–157. DOI: 10.18653/v1/2022.iwslt-1.10.

Barbaresi, Adrien (2023). *Simplemma*. Version v0.9.1. DOI: 10.5281/zenodo.7555188.

Bradski, G. (2000). "The OpenCV Library". In: *Dr. Dobb's Journal of Software Tools*.

Camgöz, Necati Cihan, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden (2018). "Neural sign language translation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7784–7793. DOI: 10.1109/CVPR.2018.00812.

Cao, Z., G. Hidalgo, T. Simon, S. Wei, and Y. Sheikh (2021). "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.01, pp. 172–186. ISSN: 1939-3539. DOI: 10.1109/TPAMI.2019.2929257.

Carreira, João and Andrew Zisserman (2017). "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4724–4733. DOI: 10.1109/CVPR.2017.502.

Chan, Caroline, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros (2019). "Everybody dance now". In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5932–5941. DOI: 10.1109/ICCV.2019.00603.

Egea Gómez, Santiago, Euan McGill, and Horacio Saggion (2021). "Syntax-aware Transformers for Neural Machine Translation: The Case of Text to Sign Gloss Translation". In: *Proceedings of the 14th Workshop on Building and Using Comparable Corpora (BUCC 2021)*. Online (Virtual Mode): INCOMA Ltd., pp. 18–27. URL: https://aclanthology.org/2021.bucc-1.4.

Erjavec, Tomaž, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubešić, Kiril Simov, Andrej Pančur, Michał Rudolf, Matyáš Kopp, Starkaður Barkarson, Steinþór Steingrímsson, Çağrı Çöltekin, Jesse de Does, Katrien Depuydt, Tommaso Agnoloni, Giulia Venturi, María Calzada Pérez, Luciana D. de Macedo, Costanza Navarretta, Giancarlo Luxardo, Matthew Coole, Paul Rayson, Vaidas Morkevičius, Tomas Krilavičius, Roberts Darģis, Orsolya Ring, Ruben van Heusden, Maarten Marx, and Darja Fišer (2021). *Linguistically annotated multilingual com-*

parable corpora of parliamentary debates ParlaMint.ana 2.1. Slovenian language resource repository CLARIN.SI. URL: http://hdl.handle.net/11356/1431.

Etchegoyhen, Thierry and Harritxu Gete (2020). "Handle with Care: A Case Study in Comparable Corpora Exploitation for Neural Machine Translation". In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 3799–3807. ISBN: 979-10-95546-34-4. URL: https://aclanthology.org/2020.lrec-1.469.

Federmann, Christian (2018). "Appraise Evaluation Framework for Machine Translation". In: *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*. Santa Fe, New Mexico: Association for Computational Linguistics, pp. 86–88. URL: https://www.aclweb.org/anthology/C18-2019.

Freitag, Markus, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey (2021). "Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation". In: *Transactions of the Association for Computational Linguistics* 9, pp. 1460–1474. DOI: 10.1162/tacl_a_00437.

Graham, Yvette, Timothy Baldwin, Meghan Dowling, Maria Eskevich, Teresa Lynn, and Lamia Tounsi (2016). "Is all that Glitters in Machine Translation Quality Estimation really Gold?" In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. Osaka, Japan: The COLING 2016 Organizing Committee, pp. 3124–3134. URL: https://aclanthology.org/C16-1294.

Grishchenko, Ivan and Valentin Bazarevsky (2020). *MediaPipe Holistic*. URL: https://google.github.io/mediapipe/solutions/holistic.html.

Grundkiewicz, Roman, Marcin Junczys-Dowmunt, Christian Federmann, and Tom Kocmi (2021). "On User Interfaces for Large-Scale Document-Level Human Evaluation of Machine Translation Outputs". In: *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*. Online: Association for Computational Linguistics, pp. 97–106. URL: https://aclanthology.org/2021.humeval-1.11.

Güler, Rıza Alp, Natalia Neverova, and Iasonas Kokkinos (2018). "Densepose: Dense human pose estimation in the wild". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7297–7306. DOI: 10.1109/CVPR.2018.00762.

Hanke, Thomas, Susanne König, Reiner Konrad, Gabriele Langer, Patricia Barbeito Rey-Geißler, Dolly Blanck, Stefan Goldschmidt, Ilona Hofmann, Sung-Eun Hong, Olga Jeziorski, Thimo Kleyboldt, Lutz König, Silke Matthes, Rie Nishio, Christian Rathmann, Uta Salden, Sven Wagner, and Satu Worseck (2020). *MEINE DGS. Öffentliches Korpus der Deutschen Gebärdensprache, 3. Release*. languageresource. Version 3.0. DOI: 10.25592/dgs.meinedgs-3.0.

Hieber, Felix, Michael Denkowski, Tobias Domhan, Barbara Darques Barros, Celina Dong Ye, Xing Niu, Cuong Hoang, Ke Tran, Benjamin Hsu, Maria Nadejde, Surafel Lakew, Prashant Mathur, Anna Currey, and Marcello Federico (2022). *Sockeye 3: Fast Neural Machine Translation with PyTorch*. DOI: 10.48550/ARXIV.2207.05851.

Isola, Phillip, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros (2017). "Image-to-image translation with conditional adversarial networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5967–5976. DOI: 10.1109/CVPR.2017.632.

Johnson, Melvin, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean (2017). "Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation". In: *Transactions of the Association for Computational Linguistics* 5, pp. 339–351. DOI: 10.1162/tacl_a_00065.

Funded by the Horizon 2020
Framework Programme of the European Union

Jwalapuram, Prathyusha, Shafiq Joty, and Youlin Shen (2020). “Pronoun-Targeted Fine-tuning for NMT with Hybrid Losses”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 2267–2279. DOI: 10.18653/v1/2020.emnlp-main.177.

El-Kishky, Ahmed, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn (2020). “CCAligned: A Massive Collection of Cross-Lingual Web-Document Pairs”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 5960–5969. DOI: 10.18653/v1/2020.emnlp-main.480.

Koehn, Philipp (2005). “Europarl: A Parallel Corpus for Statistical Machine Translation”. In: *Proceedings of Machine Translation Summit X: Papers*. Phuket, Thailand, pp. 79–86. URL: https://aclanthology.org/2005.mtsummit-papers.11.

Kudo, Taku (2018). “Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 66–75. DOI: 10.18653/v1/P18-1007.

Lebert, Marie (2008). *Project Gutenberg (1971-2008)*.

Loáiciga, Sharid, Sara Stymne, Preslav Nakov, Christian Hardmeier, Jörg Tiedemann, Mauro Cettolo, and Yannick Versley (2017). “Findings of the 2017 DiscoMT Shared Task on Cross-lingual Pronoun Prediction”. In: *Proceedings of the Third Workshop on Discourse in Machine Translation*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 1–16. DOI: 10.18653/v1/W17-4801.

Lu, Tianshuai, Noëmi Aepli, and Annette Rios (2023). “Reducing Gender Bias in NMT with FUDGE”. In: *Proceedings of the 1st Workshop on Gender-Inclusive Translation Technologies (GITT)*. Tampere, Finnland: Association for Computational Linguistics, pp. 61–69. URL: https://drive.google.com/file/d/19fNgcRZ0Qo4EZkHwXM3tG5GSdqJpY-uA/view.

Lugaresi, Camillo, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann (2019). “MediaPipe: A Framework for Building Perception Pipelines”. In: *CoRR* abs/1906.08172. arXiv: 1906.08172.

Min, Jianyuan and Jinxiang Chai (2012). “Motion Graphs++: A Compact Generative Model for Semantic Motion Analysis and Synthesis”. In: *ACM Transactions on Graphics* 31.6. ISSN: 0730-0301. DOI: 10.1145/2366145.2366172.

Montani, Ines, Matthew Honnibal, Matthew Honnibal, Sofie Van Landeghem, Adriane Boyd, Henning Peters, Paul O’Leary McCann, jim geovedi, Jim O’Regan, Maxim Samsonov, György Orosz, Daniël de Kok, Duygu Altinok, Søren Lind Kristiansen, Madeesh Kannan, Raphaël Bournhonesque, Lj Miranda, Peter Baumgartner, Edward, Explosion Bot, Richard Hudson, Raphael Mitsch, Roman, Leander Fiedler, Ryn Daniels, Wannaphong Phatthiyaphaibun, Grégory Howard, Yohei Tamura, and Sam Bozek (2023). *explosion/spaCy: v3.5.0: New CLI commands, language updates, bug fixes and much more*. Version v3.5.0. DOI: 10.5281/zenodo.7553910.

Moryossef, Amit (2023). *sign.mt: A Web-Based Application for Real-Time Multilingual Sign Language Translation*. https://sign.mt/.

Moryossef, Amit and Mathias Müller (2021). *pose-format: Library for viewing, augmenting, and handling .pose files*. https://github.com/sign-language-processing/pose.

Moryossef, Amit, Mathias Müller, Anne Göhring, Zifan Jiang, Yoav Goldberg, and Sarah Ebling (2023). “An Open-Source Gloss-Based Baseline for Spoken to Signed Language Translation”. In: *2nd International Workshop on Automatic Translation for Signed and Spoken*

Languages (AT4SSL). Available at: https://arxiv.org/abs/2305.17714. URL: https://github.com/ZurichNLP/spoken-to-signed-translation.

Moryossef, Amit, Ioannis Tsochantaridis, Joe Dinn, Necati Cihan Camgöz, Richard Bowden, Tao Jiang, Annette Rios, Mathias Müller, and Sarah Ebling (2021a). "Evaluating the Immediate Applicability of Pose Estimation for Sign Language Recognition". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. DOI: 10.1109/CVPRW53098.2021.00382.

Moryossef, Amit, Kayo Yin, Graham Neubig, and Yoav Goldberg (2021b). "Data Augmentation for Sign Language Gloss Translation". In: *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*. Virtual: Association for Machine Translation in the Americas, pp. 1–11. URL: https://aclanthology.org/2021.mtsummit-at4ssl.1.

Müller, Mathias, Sarah Ebling, Eleftherios Avramidis, Alessia Battisti, Michèle Berger, Richard Bowden, Annelies Braffort, Necati Cihan Camgöz, Cristina España-bonet, Roman Grundkiewicz, Zifan Jiang, Oscar Koller, Amit Moryossef, Regula Perrollaz, Sabine Reinhard, Annette Rios, Dimitar Shterionov, Sandra Sidler-miserez, and Katja Tissi (2022). "Findings of the First WMT Shared Task on Sign Language Translation (WMT-SLT22)". In: *Proceedings of the Seventh Conference on Machine Translation (WMT)*. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, pp. 744–772. URL: https://aclanthology.org/2022.wmt-1.71.

Müller, Mathias, Zifan Jiang, Amit Moryossef, Annette Rios, and Sarah Ebling (2023). "Considerations for meaningful sign language machine translation based on glosses". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Toronto, Canada: Association for Computational Linguistics, pp. 682–693. URL: https://aclanthology.org/2023.acl-short.60.

Othman, Achraf and Mohamed Jemni (2012). "English-ASL Gloss Parallel Corpus 2012: ASLG-PC12". In: *8th International Conference on Language Resources and Evaluation (LREC 2012). Proceedings of the LREC2012 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon*. Ed. by Onno Crasborn, Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, Jette Kristoffersen, and Johanna Mesch. Istanbul, Turkey: European Language Resources Association (ELRA), pp. 151–154. URL: https://www.sign-lang.uni-hamburg.de/lrec/pub/12019.pdf.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu (2002). "Bleu: a Method for Automatic Evaluation of Machine Translation". In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, pp. 311–318. DOI: 10.3115/1073083.1073135.

Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala (2019). "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Curran Associates, Inc., pp. 8024–8035. URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Pizzuto, Elena Antinoro, Paolo Rossini, and Tommaso Russo (2006). "Representing Signed Languages in Written Form: Questions that Need to be Posed". In: *5th International Conference on Language Resources and Evaluation (LREC 2006). Proceedings of the LREC2006 2nd Workshop on the Representation and Processing of Sign Languages: Lexicographic*

Matters and Didactic Scenarios. Ed. by Chiara Vettori. Genoa, Italy: European Language Resources Association (ELRA), pp. 1–6. URL: https://www.sign-lang.uni-hamburg.de/lrec/pub/06001.pdf.

Popović, Maja (2016). "chrF deconstructed: beta parameters and n-gram weights". In: *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*. Berlin, Germany: Association for Computational Linguistics, pp. 499–504. DOI: 10.18653/v1/W16-2341.

Post, Matt (2018). "A Call for Clarity in Reporting BLEU Scores". In: *Proceedings of the Third Conference on Machine Translation: Research Papers*. Brussels, Belgium: Association for Computational Linguistics, pp. 186–191. DOI: 10.18653/v1/W18-6319.

Prillwitz, Siegmund and Heiko Zienert (1990). "Hamburg Notation System for Sign Language: Development of a sign writing with computer application". In: *Current trends in European Sign Language Research. Proceedings of the 3rd European Congress on Sign Language Research*, pp. 355–379.

Rei, Ricardo, Craig Stewart, Ana C Farinha, and Alon Lavie (2020). "COMET: A Neural Framework for MT Evaluation". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 2685–2702. DOI: 10.18653/v1/2020.emnlp-main.213.

Saunders, Ben, Necati Cihan Camgöz, and Richard Bowden (2020). "Everybody Sign Now: Translating Spoken Language to Photo Realistic Sign Language Video". In: *arXiv preprint arXiv:2011.09846*.

— (2021). "Anonysign: Novel Human Appearance Synthesis for Sign Language Video Anonymisation". In: *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pp. 1–8. DOI: 10.1109/FG52635.2021.9666984.

Savitzky, Abraham and Marcel JE Golay (1964). "Smoothing and differentiation of data by simplified least squares procedures." In: *Analytical chemistry* 36.8, pp. 1627–1639.

Savoldi, Beatrice, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi (2022). "Under the Morphosyntactic Lens: A Multifaceted Evaluation of Gender Bias in Speech Translation". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 1807–1824. DOI: 10.18653/v1/2022.acl-long.127.

Schweizerischer Gehörlosenbund SGB-FSS (2023). *Gehörlosenbund Gebärdensprache-Lexikon*. https://signsuisse.sgb-fss.ch/. Accessed on: July 27, 2023.

Shalev-Arkushin, Rotem, Amit Moryossef, and Ohad Fried (2023). "Ham2Pose: Animating Sign Language Notation into Pose Sequences". In: *CoRR* abs/2211.13613. arXiv: 2211.13613 [cs.CV].

Shieber, Stuart (1994). "Restricting the weak-generative capacity of synchronous tree-adjoining grammars". In: *Computational Intelligence* 10.4, pp. 371–385.

Shieber, Stuart M. and Yves Schabes (1990). "Synchronous Tree-Adjoining Grammars". In: *COLING 1990 Volume 3: Papers presented to the 13th International Conference on Computational Linguistics*. URL: https://aclanthology.org/C90-3045.

Stanovsky, Gabriel, Noah A. Smith, and Luke Zettlemoyer (2019). "Evaluating Gender Bias in Machine Translation". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 1679–1684. DOI: 10.18653/v1/P19-1164.

Stoll, Stephanie, Necati Cihan Camgöz, Simon Hadfield, and Richard Bowden (2018). "Sign language production using neural machine translation and generative adversarial networks". In: *Proceedings of the 29th British Machine Vision Conference (BMVC 2018)*. British Machine Vision Association.

Stoll, Stephanie, Necati Cihan Camgöz, Simon Hadfield, and Richard Bowden (2020). “Text2Sign: towards sign language production using neural machine translation and generative adversarial networks”. In: *International Journal of Computer Vision*, pp. 1–18.

Sutton, Valerie (1990). *Lessons in sign writing*. SignWriting.

Tarrés, Laia, Gerard I Gállego, Amanda Duarte, Jordi Torres, and Xavier Giró-i-Nieto (2023). “Sign Language Translation from Instructional Videos”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5624–5634.

Uthus, David C., Santiago Ontañón, Joshua Ainslie, and Mandy Guo (2023). “mLongT5: A Multilingual and Efficient Text-To-Text Transformer for Longer Sequences”. In: *CoRR* abs/2305.11129. DOI: 10.48550/arXiv.2305.11129. arXiv: 2305.11129.

Vanmassenhove, Eva and Christian Hardmeier (2018). “Europarl Datasets with Demographic Speaker Information”. In: *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*. Alicante, Spain, p. 391. URL: https://aclanthology.org/2018.eamt-main.59.

Vanmassenhove, Eva, Christian Hardmeier, and Andy Way (2018). “Getting Gender Right in Neural Machine Translation”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 3003–3008. DOI: 10.18653/v1/D18-1334.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). “Attention is All you Need”. In: *Advances in Neural Information Processing Systems 30*, pp. 5998–6008.

Ventura, Lucas, Amanda Cardoso Duarte, and Xavier Giró-i-Nieto (2020). “Can Everybody Sign Now? Exploring Sign Language Video Generation from 2D Poses”. In: *CoRR* abs/2012.10941. arXiv: 2012.10941.

Wang, Ting-Chun, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro (2018). “Video-to-Video Synthesis”. In: *Advances in Neural Information Processing Systems (NeurIPS)*.

Xue, Linting, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel (2021). “mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer”. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, pp. 483–498. DOI: 10.18653/v1/2021.naacl-main.41.

Yang, Kevin and Dan Klein (2021). “FUDGE: Controlled Text Generation With Future Discriminators”. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, pp. 3511–3535. DOI: 10.18653/v1/2021.naacl-main.276.

Yin, Kayo and Jesse Read (2020). “Better Sign Language Translation with STMC-Transformer”. In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, pp. 5975–5989. DOI: 10.18653/v1/2020.coling-main.525.

Zhao, Liwei, Karin Kipper, William Schuler, Christian Vogler, Norman Badler, and Martha Palmer (2000). “A machine translation system from English to American Sign Language”. In: *Conference of the Association for Machine Translation in the Americas*. Springer, pp. 54–67.

5 TRANSLATION: INSTRUCTIONS TO HUMAN EVALUATORS

5.1 SIGN-TO-SPOKEN EVALUATION

Below you see a document with 10 sentences in Swiss-German Sign Language (Deutschschweizer Gebärdensprache (DSGS)) (left columns) and their corresponding candidate translations in German (Deutsch) (right columns). Score each candidate sentence translation in the document context. You may revisit already scored sentences and update their scores at any time by clicking at a source video.

Assess the translation quality on a continuous scale using the quality levels described as follows:

0: Nonsense/No meaning preserved: Nearly all information is lost between the translation and source. Grammar is irrelevant. 2: Some Meaning Preserved: The translation preserves some of the meaning of the source but misses significant parts. The narrative is hard to follow due to fundamental errors. Grammar may be poor. 4: Most Meaning Preserved and Few Grammar Mistakes: The translation retains most of the meaning of the source. It may have some grammar mistakes or minor contextual inconsistencies. 6: Perfect Meaning and Grammar: The meaning of the translation is completely consistent with the source and the surrounding context. The grammar is also correct.

5.2 SPOKEN-TO-SIGN EVALUATION

Below you see a document with 10 sentences in German (Deutsch) (left columns) and their corresponding candidate translations in Swiss-German Sign Language (Deutschschweizer Gebärdensprache (DSGS)) (right columns). Score each candidate sentence translation in the document context. You may revisit already scored sentences and update their scores at any time by clicking at a source text.

Assess the translation quality on a continuous scale using the quality levels described as follows:

0: Nonsense/No meaning preserved: Nearly all information is lost between the translation and source. Naturalness of motion is irrelevant. 2: Some Meaning Preserved: The translation preserves some of the meaning of the source but misses significant parts. The narrative is hard to follow due to fundamental errors. Naturalness of motion may be poor. 4: Most Meaning Preserved and Acceptable Natural Motion: The translation retains most of the meaning of the source. It may have some minor mistakes or contextual inconsistencies. Motion may appear unnatural. 6: Perfect Meaning and Naturalness: The meaning of the translation is completely consistent with the source and the surrounding context. Motion is natural.

Funded by the Horizon 2020
Framework Programme of the European Union

5.3 ADDITIONAL RESULTS FOR SIGNED-TO-SPOKEN EXPERIMENTS

Table 5.1 shows additional scores for the signed-to-spoken translation systems, listing BLEU scores instead of CHRF.

			DSGS→DE	**LSF→FR**	**LIS→IT**	**BSL→EN**	**DGS→DE**
	(pre)trained on	**finetuned on**					
mediapipe	parallel	-	0.089	0.075	0.015	0.033	-
EMSL v1.0b DGS	parallel	-	0.394	0.252	**0.350**	-	-
EMSL v1.0b BSL	parallel	-	0.348	0.241	0.308	0.520	-
mediapipe	comparable	-	0.215	0.073	0.122	0.187	0.079
EMSL v1.0b DGS	comparable	-	0.339	0.575	0.163	-	**1.841**
EMSL v1.0b BSL	comparable	-	0.556	0.295	0.161	0.436	1.319
mediapipe	comparable	parallel	0.081	0.032	0.011	0.162	-
EMSL v1.0b DGS	comparable	parallel	0.231	0.840	0.140	-	-
EMSL v1.0b BSL	comparable	parallel	0.263	0.134	0.147	0.432	-
EMSL v2.0b DGS	parallel	-	0.210	0.505	0.132	-	-
EMSL v2.0b BSL	parallel	-	0.212	0.523	0.228	0.486	-
EMSL v2.0b BOTH	parallel	-	0.226	0.321	0.140	-	-
EMSL v2.0b DGS	both	-	**0.613**	0.348	0.081	-	-
EMSL v2.0b BSL	both	-	0.285	**0.979**	0.292	**2.888**	-
EMSL v2.0b BOTH	both	-	0.507	0.911	0.204	-	-

Table 5.1: *Translation quality measured by BLEU on the EASIER manually corrected test data.*