Findings of the Second WMT Shared Task on Sign Language Translation (WMT-SLT23)

Mathias Müller University of Zurich
Malihe Alikhani Northeastern University
Eleftherios Avramidis DFKI Berlin
Richard Bowden University of Surrey
Annelies Braffort University of Paris-Saclay
Necati Cihan Camgöz Meta Reality Labs
Sarah Ebling University of Zurich
Cristina España-Bonet DFKI Saarbrücken
Anne Göhring University of Zurich
Roman Grundkiewicz Microsoft
Mert Inan Northeastern University
Zifan Jiang University of Zurich
Oscar Koller Microsoft
Amit Moryossef Bar-Ilan University
Annette Rios University of Zurich
Dimitar Shterionov Tilburg University
Sandra Sidler-Miserez HfH Zurich
Katja Tissi HfH Zurich
Davy Van Landuyt European Union of the Deaf

Abstract

This paper presents the results of the Second WMT Shared Task on Sign Language Translation (WMT-SLT23)[1]. This shared task is concerned with automatic translation between signed and spoken[2] languages. The task is unusual in the sense that it requires processing visual information (such as video frames or human pose estimation) beyond the well-known paradigm of text-to-text machine translation (MT). The task offers four tracks involving the following languages: Swiss German Sign Language (DSGS), French Sign Language of Switzerland (LSF-CH), Italian Sign Language of Switzerland (LIS-CH), German, French and Italian. Four teams (including one working on a baseline submission) participated in this second edition of the task, all submitting to the DSGS-to-German track. Besides a system ranking and system papers describing state-of-the-art techniques, this shared task makes the following scientific contributions: novel corpora and reproducible baseline systems. Finally, the task also resulted in publicly available sets of system outputs and more human evaluation scores for sign language translation.

1 Introduction

This paper presents the outcome of the Second WMT Shared Task on Sign Language Translation (WMT-SLT23). This shared task focuses on automatic translation between signed and spoken languages. Our main goal is working towards including signed languages in NLP research (Yin et al., 2021).

Sign language translation requires processing visual information (such as video frames or human pose estimation) beyond the well-known paradigm of text-to-text machine translation (MT). As a consequence, viable solutions need to consider a combination of Natural Language Processing (NLP), computer vision (CV), computer graphics and animation techniques.

We build on and extend the work done for the first shared task on sign language translation (WMT-SLT22; Müller et al., 2022). Compared to the first edition, we

- extended our competition to more languages (three language pairs instead of one),
- provided much more training data for Swiss German Sign language compared to last year (437 hours instead of 16),
- emphasized sign languages as the target language instead of the source, for instance, by offering official baseline systems for spoken-to-signed translation (not offered last year).

In this second edition of the shared task, we considered the following languages: Swiss German Sign Language (DSGS), French Sign Language of Switzerland (LSF-CH), Italian Sign Language

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

[2] In this paper we use the word "spoken" to refer to any language that is not signed, no matter whether it is represented as text or audio, and no matter whether the discourse is formal (e.g. writing) or informal (e.g. dialogue).

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 68–94
December 6–7, 2023.

of Switzerland (LIS-CH), German, French, and Italian. We offered four tracks: DSGS-to-German translation, German-to-DSGS translation, French-to-LSF translation, and Italian-to-LIS translation.

Four teams participated in the task, which we consider a success. All teams submitted to the DSGS-to-German track, while there were no submissions to any of the tracks where a sign language is the target language.

The remainder of this paper is organized as follows:

- We give some background on sign languages and sign language processing in §2.
- We describe the shared task tracks and submission procedure in §3.
- We report on the corpora we built and distributed specifically for this task in §4 and §5.
- We describe all submitted systems, including our baselines in §6.
- We ran both an automatic and a human evaluation. We explain our evaluation in §7.
- We share the main outcomes in §8 and discuss in §9.

2 Background

In recent years, Sign Language Processing (SLP) has emerged as a sub-area of Natural Language Processing (NLP). Within this field, automatic sign language translation (SLT; or sign language machine translation, SLMT) represents a more specialized discipline, aiming to develop technology that facilitates translation between sign languages and spoken or written languages, but also between sign and sign languages. However, the challenges related to SLP and SLT differ from those of NLP and MT for spoken languages in both range and complexity. Due to the different modality, lack of structured, high-quality, high-quantity data, and the lack of NLP tools, joint efforts from the fields of sign linguistics and computational linguistics, computer science, machine learning, computer vision, 3D animation and others are needed in order to advance this field.

In this section we give an introduction to sign languages (§2.1) and describe the societal and academic relevance of SLP (§2.2). Then we give an overview of SLP in general (§2.3) and of SLT in particular (§2.4) For a general motivation for a shared task involving sign languages see Müller et al. (2022).

2.1 Sign languages

Sign languages are natural languages with their own grammatical structures and lexicons, primarily used by the deaf and hard-of-hearing communities. Contrary to the popular belief that sign language is universal, hundreds of different SLs have been documented so far.

Nature of sign languages Sign languages are visuo-gestural languages. A signer conveys an utterance using their body: through the expression of manual features (hand configuration, location, and orientation) and non-manual features (including facial expressions, mouthing and mouth gestures, gaze and torso direction). The linguistic system of SLs makes use of these specific channels. Information is expressed simultaneously (as opposed to the sequential nature of spoken language), organized in three-dimensional space, and iconicity plays a central role (Woll, 2013; Perniss et al., 2015; Slonimska et al., 2021).

Writing systems To date, SLs have no universally accepted written form or graphical system for transcription (Pizzuto and Pietrandrea, 2001; Filhol, 2020). Several notation systems, such as HamNoSys (Hanke, 2004) or SignWriting (Sutton, 1990; Bianchini and Borgia, 2012), are used in research or teaching but are rarely adopted as a writing system in everyday life, limiting the standardisation of data collection and processing. In SL research, a common practice is therefore to use glosses – text-based, semantic labels for signs, typically borrowed from the corresponding regional spoken language.

A common misconception among MT researchers is that transcribed glosses are a full-fledged writing system for sign languages. In reality, glossing can only be seen a linguistic tool, useful for annotating corpora for linguistic studies (Johnston, 2010). Glosses do not adequately represent the meaning of an SL utterance and, more importantly, "deaf people do not read or write glosses" in everyday life (Müller et al., 2023).

2.2 Relevance of sign language processing

SLP is a research area with high potential societal and academic impact.

Societal impact The overall aim of SLP is to provide language technology for sign languages, which currently are somewhat overlooked, since the vast majority of NLP systems are designed only for spoken languages. This means that more research in SLP could result in more equal access to language technology.

The more specific goal of SLT is to facilitate communication between the deaf and hard-of-hearing communities on the one side and the hearing community on the other side. There is a need for this because speakers of spoken languages and signers of sign languages experience communication difficulties (the same kind of difficulties encountered by speakers of different spoken languages). We emphasize that these technologies should be developed in such a way, so that deaf/hard-of-hearing and hearing people can benefit from them in an equal measure.[3]

Besides aiding direct communication, SLT would improve accessibility to spoken language content, given that spoken languages are often a second language for deaf people, where they exhibit varying proficiency. The reverse direction is also crucial, for example to automatically subtitle signed content to make it accessible to people who do not know sign languages (Bragg et al., 2019).

Academic relevance In the field of NLP, working on sign languages is highly innovative and timely. Recently, a call for more inclusion of signed languages in NLP (Yin et al., 2021) was widely publicized, and an ACL initiative for Diversity and Inclusion[4] targets SL processing as well.

And even though sign languages are still a niche topic in the general field of NLP (the vast majority of NLP systems are designed for spoken languages, not for signed languages), the advancement and spread of SLP tools, calls, initiatives and events lead to knowledge transfer not only within the academic spheres, or between researchers, developers and users, but also, more importantly, between deaf, hard-of-hearing and hearing individuals involved in the process.

2.3 Sign language processing

Sign language processing is an interdisciplinary field, bringing together research on NLP and computer vision, among other disciplines (Bragg et al., 2019). For a general overview in the context of NLP see Yin et al. (2021); Moryossef and Goldberg (2021).

Tasks SLP involves a variety of (sub)tasks with individual challenges. Widely known tasks are sign language recognition, sign language translation, and sign language production (or *synthesis*). Sign language recognition usually refers to identifying individual signs from videos; see Koller (2020) for an overview. Sign language translation refers to the task of transforming sign language data to a second language, no matter whether signed or spoken; see De Coster et al. (2022) for a comprehensive survey. Finally, sign language production refers to rendering sign language as a video, using methods such as avatar animation (Wolfe et al., 2022) or video generation.

SLP research is challenging for a number of different reasons. The ones we chose to highlight here are linguistic properties, availability of data, and availability of basic NLP tools.

Linguistic challenges SLP is challenging because the characteristics of sign languages (§2.1) cannot be fully handled with existing methods, for instance, the multilinearity, the use of the signing space, and the iconicity. As explained earlier, SLP needs to take into account manual and non-manual cues in order to capture a complete linguistic picture of an SL utterance (Crasborn, 2006). Information is spatio-temporal in nature and the data is simultaneously conveyed by a number of articulators. Signing makes frequent use of indexing strategies for example to identify referents introduced earlier in the discourse or timelines (Engberg-Pedersen, 1993). In other words, a sign language utterance is not a simple sequence of lexical units.

Sign languages have an established vocabulary but are also lexically productive to allow for the definition of new signs or constructions to be used to depict entities or situations (Johnston, 2011).

Availability of data Given the current research landscape in NLP, sign languages are under-resourced. An analysis by Joshi et al. (2020) places all sign languages considered in this study in the category "left behind" (together with many spoken

[3] We distance ourselves from the audistic view that only deaf people are in need (of access to spoken language discourse). Language barriers are inherently two-way, and addressing them involves both parties.

[4] `https://www.2022.aclweb.org/dispecialinitiative`

languages). Existing resources are small and heterogeneous. They are created under a variety of circumstances and vary in quality (e.g. video resolution), signer demographics (e.g. deaf vs. hearing signers), richness of annotation (e.g. glosses, sentence segmentation, translation to a spoken language), and linguistic domain (e.g. only weather reports, hence a very limited domain).

Also, not all corpora are easily accessible online and some have restrictive licenses that disallow NLP research. A survey of SL corpora available in Europe can be found in Kopf et al. (2021). For an account of further challenges relating to data see De Sisto et al. (2022).

Lack of basic linguistic tools SLP currently lacks fundamental NLP tools that are readily available for spoken languages. Such tools include automatic language identification (Monteiro et al., 2016), sign segmentation (De Sisto et al., 2021), sentence segmentation (Ormel and Crasborn, 2012; Bull et al., 2020b) and sentence alignment (Varol et al., 2021). Although there are experimental solutions, they are not yet viable in practice.

Tools like these would be crucial to create better corpora by constructing them automatically, as is routinely done for spoken languages (Bañón et al., 2020), and develop better high-level NLP solutions.

2.4 Sign language translation

In recent years, different methods to tackle SLT have been proposed, most of them suggesting a cascaded system where a signed video is first converted to an intermediate representation and then to spoken text (similarly for text-to-video translation). Intermediate representations (with individual strengths and weaknesses) include pose estimation (§5.3), glosses or writing systems such as HamNoSys (§2.1, writing systems).

There is existing work on gloss-to-text translation (e.g. Camgöz et al. 2018; Yin and Read 2020) and vice versa (e.g. Stoll et al., 2020), pose-to-text translation and vice versa (e.g. Ko et al. 2019; Saunders et al. 2020a,b,c; Inan et al. 2022; Viegas et al. 2023) and systems involving HamNoSys (e.g. Morrissey 2011; Walsh et al. 2022), or AZee expressions, designed to be used as input to avatar synthesis systems (Bertin-Lemée et al., 2023). Recently, direct video-to-text translation was also proposed by Camgöz et al. (2020a,b). For rendering sign language output, avatars are commonly used (Wolfe et al., 2022), as well as methods to generate videos of realistic signers (e.g. Saunders et al. 2022).

Parallel datasets In terms of datasets, past work in SLT can be characterized as focusing very much on a narrow linguistic domain, most of the work was done on one single data set called `RWTH-PHOENIX Weather 2014T` (Forster et al., 2014). PHOENIX has a size of 8k sentence pairs and contains only weather reports. The biggest parallel corpus for a European sign language to date, the Public DGS Corpus (Hanke et al., 2020), contains roughly 70k sentence pairs.

Thus, there is a clear shortage of usable parallel corpora, and existing ones are orders of magnitude smaller than what is considered an acceptable size for spoken language MT (as a rule of thumb, at least hundreds of thousands of sentence pairs). Nevertheless, there are plenty of spoken languages that also have little parallel data and MT methods have been developed specifically for low-resource MT (Sennrich and Zhang, 2019).

Evaluation For spoken language MT a variety of automatic metrics exist. These include more conventional, string-based metrics such as BLEU (Papineni et al., 2002) or chrF (Popović, 2015), as well as recent, learned metrics based on embeddings like COMET (Rei et al., 2020). In the context of SLT, no automatic metrics are validated empirically, but if the target language is spoken, many existing metrics are reasonable to use. However, if sign language is the target language, no automatic metric is known at the time of writing, and the only viable evaluation method is human evaluation. Apart from last year's shared task, a human evaluation of SLT systems has never been conducted on a large scale before, and there are open questions regarding the exact evaluation methodology and what the ideal profile (e.g. hearing status, language proficiency) for evaluators should be.

3 Tracks and submission procedure

We offered four translation directions ("tracks"): translation from DSGS to German and vice versa, French to LSF-CH, and Italian to LIS-CH.

For DSGS to German, submitted systems were ranked on a leaderboard. For all other directions, no automatic ranking was shown since automatic metrics of translation quality do not exist for sign languages as the target language.

We provided baseline systems for both translation scenarios (translating from or to a sign language). We were prepared to provide human evaluation for all submitted systems, regardless of the translation direction or language pair.

We deliberately did not limit the shared task to any particular kind of SL representation as input or output of an MT system. For DSGS-to-German translation, participants were free to use video frames, pose estimation, or something else. For German-to-DSGS participants were free to submit a video showing pose estimation output, an avatar, or a photo-realistic signer.

Participants had to submit their translation outputs on the OCELoT platform[5] which displayed an unofficial public leaderboard based on automatic metrics. Participants were allowed to make up to seven submissions and were asked to mark one of them as their primary submission.

Main outcome Four teams (including one from Northeastern University whose submission we consider a baseline) participated in our task. All of them submitted to the DSGS-to-German track, while there were no submissions for other translation directions.

4 Data

For this task we provided separate training, development and test data. While the training data was available from the beginning, the test data has been released in two stages, starting with a release of the test sources only.

Table 1 gives a high-level overview of our training, development and test data.

4.1 Licensing and attribution

Both datasets (SRF23 and Signsuisse) can be used for non-commercial research. Please note that distributing the datasets or making them accessible to third parties is not permitted, either in their original or edited form. In addition, this overview paper should be cited if the corpora are used.

4.2 Training Data

The training data comprises two corpora called Signsuisse (Jiang et al., 2023a) and SRF23 (Jiang et al., 2023b). Signsuisse is a multilingual dictionary containing lexical items in DSGS, LSF-CH and LIS-CH, represented as videos and glosses. Additionally, Signsuisse contains sentence-level parallel data as well, since there is one example sentence to show the use of the sign in context for each lexical item. SRF23 contains parallel data between DSGS and German, and its linguistic domain is general news. Both datasets are distributed through SwissUbase[6], where individual researchers had to agree with the usage terms and apply for access before downloading.

Training corpus 1: Signsuisse Lexicon We collected $18,221$ lexical items from the Signsuisse website, $17,221$ of which are released as training data and $1,000$ are reserved for testing and therefore not included in the training data release. The lexicon contains three languages: (i) DSGS (9044 items, 500 reserved), (ii) LSF-CH (6423 items, 250 reserved), and (iii) LIS-CH (2754 items, 250 reserved).

The lexical items are represented as videos and glosses, which enable sign-by-sign translation from spoken to signed languages. The videos were recorded with different framerates, either 24, 25, or 30 fps, and the video resolution is 640 x 480.

Training corpus 2: SRF23 These are daily national news and weather forecast episodes broadcast by the Swiss National TV (Schweizerisches Radio und Fernsehen, SRF)[7]. The episodes are narrated in Standard German of Switzerland (different from Standard German of Germany, and different from Swiss German dialects) and interpreted into Swiss German Sign Language (DSGS). The interpreters are hearing individuals, some of them children of Deaf adults (CODAs).

The subtitles are partly preproduced, and partly created live via respeaking to automatic speech recognition. 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 (Müller et al., 2022). It should also be pointed out that there are differences between interpreted and non-interpreted language (Dayter, 2019) due to source language interference and time constraints. SL during real-time interpretation tends to closely follow the grammatical structure of the spoken language (Leeson, 2005).

[5]https://ocelot-wmt23.mteval.org/

[6]https://www.swissubase.ch/en/catalogue/studies/20452/19280/overview

[7]https://www.srf.ch

	direction	SRF23 episodes	SRF23 segments	Signsuisse segments	Signsuisse lexical items	Total segments	Total lexical items
training	DSGS↔DE	771	231834	9044	9044	240878	9044
	FR→LSF-CH	-	-	6423	6423	6423	6423
	IT→LIS-CH	-	-	2754	2754	2754	2754
development	DSGS↔DE	3	712	-	-	712	-
test	DSGS→DE	1	246	250	250	496	250
	DE→DSGS	1	258	250	250	508	250
	FR→LSF-CH	-	-	250	250	250	250
	IT→LIS-CH	-	-	250	250	250	250

Table 1: Overview of training, development and test data. SRF23 and Signsuisse are two different training corpora (§4.2). Segment count for the training corpora is after automatic sentence segmentation. The training data and development data for DSGS→DE and DE→DSGS are identical, while the test data is different. There was no designated development data for LSF-CH and LIS-CH.

Different from the first edition of the shared task (WMT-SLT22), the offset between the signing and the subtitles was not manually corrected for the training data of the current edition. On the other hand, the size of the training data is much larger than last year, presenting a different trade-off. See Table 2 for a comparison between this year's and last year's SRF resources. While last year our focus was providing training data of the highest quality, this year our focus was offering a large, noisy dataset that lends itself to data cleaning or filtering experiments such as automatic alignment.

Additional resources We encouraged participants to consider the MEDIAPI-SKEL corpus with parallel examples between French Sign Language and French (Bull et al., 2020a) as a further resource. Besides, we suggested that participants re-use the training corpora released for last year's shared task (Müller et al., 2022).

4.3 Development data

We did not provide any dedicated development data for this edition of the shared task. As is customary for WMT shared tasks, we encouraged participants to use last year's development and test data as development data for the current year.

4.4 Test data

We distribute separate test data for our four translation directions. See Table 1 for an overview.

DSGS→DE The test data consists of segments taken from undisclosed SRF23 and Signsuisse material (see §4.2 for a general description). The final test set is balanced, containing roughly 50% Signsuisse and 50% SRF23 examples. For the SRF23 part one episode was manually aligned using the iLex editor (Hanke and Storz, 2008), and the signer is a "known" person that appeared in the training set. We did not intend to test generalization to unknown signers during the shared task evaluation campaign. For the Signsuisse part we do not use the isolated lexical entries themselves for testing, but the example sentences associated with each lexical item.

DE→DSGS Same procedure as DSGS→DE, except that a different SRF23 episode and different sentences from Signsuisse are reserved for this translation direction.

FR→LSF-CH 250 undisclosed sentences from Signsuisse.

IT→LIS-CH 250 undisclosed sentences from Signsuisse.

5 Data preprocessing

For each data set described in §4 we provided videos and corresponding text in a spoken language. In addition, we included pose estimates (location of body keypoints in each frame) as a convenience.

5.1 Video processing (only SRF23)

Videos are re-encoded with lossless H264 and use an mp4 container. The framerate of videos is unchanged, meaning either 25, 30 or 50. We are not distributing the original videos but ones that are preprocessed in a particular way so that they only show the part of each frame where the signer is located (cropping) and the background is replaced with a monochrome color (signer masking), see Figure 1 for examples.

	SRF22	SRF23
Number of episodes	29	771
Time span of episodes	March 2020 to March 2021	July 2014 to May 2021
Total duration videos	16 hours	437 hours
Total number of subtitles (before/after sentence segmentation)	14265 / 7071	354901 / 231834
Number of signers	3	4
Subtitle segmentation	manual	automatic
Subtitle alignment	manual	audio

Table 2: Comparison between SRF training data of the 2022 and 2023 edition of the WMT-SLT shared task. Subtitle segmentation=ensuring that each subtitle unit is one entire sentence. Subtitle alignment=Subtitle times are either manually corrected to match the signing in the video (manual) or are matched with the audio track (audio).

Figure 1: Illustration of video preprocessing steps (cropping, instance segmentation and masking). From left to right: original frame, cropped frame, masked frame. Taken from Müller et al. (2022).

Cropping We manually annotate a rectangle (bounding box) around where the signer is located for each video. We then crop the video to only keep this region using the FFMPEG library.

Signer segmentation and masking To the cropped video we apply an instance segmentation model, Solo V2 (Wang et al., 2020), to separate the background from the signer. This produces a mask that can be superimposed on the cropped video to replace each background pixel in a frame with a grey color (`[127,127,127]` in RGB).

The video processing steps described above are only necessary for the SRF23 data, since Signsuisse footage is recorded against a neutral background and showing only one signer in the center of each frame.

5.2 Subtitle processing (only SRF23)

Since SRF23 subtitles are not manually aligned, automatic sentence segmentation[8] is used to redistribute text across subtitle segments, see Table 3 for examples. This process also adjusts timecodes in a heuristic manner if needed. For instance, if automatic sentence segmentation detects that a well-formed sentence stops in the middle of a subtitle, a new end time will be computed. The end time is proportional to the location of the last character of the sentence, relative to the entire length of the subtitle. See Example 2 in Table 3 for an illustration of this case.

5.3 Pose processing (both corpora)

"Poses" are an estimate of the location of body keypoints in video frames. The exact set of keypoints depends on the pose estimation system, well-known ones are OpenPose (Cao et al., 2019)[9] and MediaPipe Holistic (Lugaresi et al., 2019)[10]. Usually such a system provides 2D or 3D coordinates of keypoints in each frame, plus a confidence value for each keypoint.

The input for pose processing are cropped and masked videos (§5.1). See Figure 2 for examples of pose estimation on our data.

OpenPose We use the Openpose 137 model (which is the default) for the Signsuisse data and the Openpose 135 model for the SRF data. The two models are both widely used and the 137 model has two additional keypoints because it represents

[8] `https://github.com/bricksdont/srt/tree/sentence_segmentation`

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

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

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.

Table 3: Examples of automatic sentence segmentation for German subtitles. The subtitles are formatted as SRT, a common subtitle format. Taken from Müller et al. (2022).

Figure 2: Examples of the output of pose estimation systems overlaid over the original video frames. Left: OpenPose, right: MediaPipe Holistic. Taken from Müller et al. (2022).

the wrists twice. OpenPose often detects several people in our videos, even though there is only one single person present. We distribute the original predictions which contain all people that OpenPose detected.

MediaPipe Holistic As an alternative, we also estimate signers' poses with the MediaPipe Holistic system developed by Google. Unlike our OpenPose model, which only provides 2D joint locations, MediaPipe produces both 2D and 3D joint location coordinates. For the SRF data, values from Holistic are normalized between 0 and 1, instead of referring to actual video coordinates.

Unlike the first edition of the task, where the keypoints were stored in a JSON format, to deliver the pose data for more compact storage and faster I/O, in WMT-SLT 23 the binary .pose format of Moryossef and Müller (2021) was used.

6 Baselines and submitted systems

In this section we describe the submissions to our shared task. In case there are substantial differences between the primary and secondary submissions of a team we opted to describe the primary submission here. At the time of writing this overview paper three out of four teams have given us detailed information about their submissions. The submissions are summarized in Table 4.

Overall, the participating teams have diverse academic backgrounds, but their expertise is leaning towards NLP more than computer vision. All submitted systems are sequence-to-sequence models based on Transformers (Vaswani et al., 2017). Participants mostly chose to represent sign language data as video frames (using a visual feature extractor on the encoder side). Only the baseline system opted for Mediapipe pose features instead.

Two systems, by KNOWCOMP and TTIC, are unconstrained because their visual or spoken text components are pretrained on other datasets. Their approaches are best summarized as a combination of visual embeddings and pre-trained language models. TTIC used additional monolingual video data from OpenASL for pretraining, and no submission used monolingual text in a spoken language.

Two teams have published their code, with another team planning to do so in the future.

6.1 Baseline by Northeastern University (DSGS→DE)

Based on the models of the previous challenge, we pre-train the baseline signed-to-spoken system using a Transformer architecture. We use the fairseq seq2seq translation library (Ott et al., 2019), and the open-source implementation of the architecture by Tarrés et al. (2023). We first train a Sentencepiece tokenization model on the German text of the example sentences of the Signsuisse dataset. Then, we train the model on the Mediapipe Holistic poses on the Signsuisse example sentences. We, then, validate and test the model on the extracted Mediapipe Holistic poses of both the Signsuisse and SRF DSGS-to-German datasets. The final output is detokenized to result in spoken German text.

6.2 Baseline by UZH (DE→DSGS, FR→LSF-CH, IT→LIS-CH)

As a naive solution, we choose a sign-by-sign translation baseline (Moryossef et al., 2023). The system gets German text as input, performs text-to-gloss translation, then for each gloss looks up a sign in the Signsuisse lexicon. The estimated poses from each sign are then concatenated and smoothed out, to create a single pose video with the translation into a sign language.

Since there were no submissions by participants to these tracks, this baseline was not used for any subsequent evaluation.

6.3 Submission by KNOWCOMP (Xu et al., 2023)

The team proposed a framework which combines a pre-trained visual model to extract visual embeddings with a GPT2-based language model to translate into text.

The framework first utilises an I3D model (Varol et al., 2022) pre-trained on the BSL-1K corpus (Albanie et al., 2020) to extract 1024-dimensional tensors for a 64-frame video input. The video extractor, i.e. the I3D model, generates a 1024-dimensional tensors as the visual representation of the input video (64 frames). For decoding, a German-GPT2 (Radford et al., 2019) large language model (LLM) is used to generate the final translations. To establish an alignment between the visual and the textual embeddings from the two models, the team trains an embedding alignment block to project the obtained visual embeddings into textual embeddings.

	BASELINE	KNOWCOMP	TTIC	CASIA
Constrained	✔	-	-	?
Multilingual	-	-	-	?
Document-level	-	-	-	?
Model ensemble	-	-	-	?
Pretrained components	-	✔	✔	?
Monolingual data	-	✔	✔	?
Synthetic data	-	-	-	?
Signed language representation	Mediapipe	I3D features	Video frames	?
Spoken language representation	SP	BPE	SP	?
Open-source code	✔	(✔)	✔	?

Table 4: Overview of characteristics of submitted systems. CASIA did not disclose any information. In the code row, checkmarks are clickable links. BPE=Byte Pair Encoding, SP=Sentencepiece, (✔)=authors plan to publish the code.

This is implemented by stacking 6 Transformer encoder layers together. Two fully connected neural networks are placed before and after the alignment block to extend the visual embeddings into a sequential format and to densify the aligned embeddings into prefix embeddings for German-GPT2, respectively.

Before training their model KnowComp first employs a data preprocessing step where the raw data is divided into smaller video segments which are then matched with the corresponding ground truth German translations. To ensure that the input observes the visual model requirements, i.e. input of 64 frames, they downsample the video segments taking the first of each three frames. In cases where the video segment is smaller than 64 frames, pure black frames are appended. Next, the video frames are resized to 224 x 224.

At training time, to enhance training efficiency, the parameters of the visual and the translation models are first frozen; later, at a certain iteration, the parameters of GPT2 are unfrozen. This strategy ensures that the randomly initialized Transformer encoder does not compromise the LLM. The hyperparameters they used are: batch size of 4, learning optimizer Adam (Kingma and Ba, 2015) with a learning rate of $5e-6$, and unfreezing the training parameters at iteration 66000. The input and output lengths of GPT2 were set to 20. The number of heads in the multi-head attention was set to 8; the prefix length for GPT2 to 4. Before the visual embeddings were fed to the alignment block, the sequence length was adjusted to 2×4, where 4 is the GPT2's prefix number. They ran their experiments on an NVIDIA GeForce GTX 1080 Ti with 11G VRAM.

6.4 Submission by TTIC (Sandoval-Castaneda et al., 2023)

The system by the TTIC team uses as visual backbone the VideoSwin Transformer (Liu et al., 2022) and the T5 model by Raffel et al. (2020) for translation into text. The VideoSwin model was pretrained on the visual (video) side of OpenASL (Shi et al., 2022, thus excluding the English translations) using the codebook from a discrete variational auto-encoder (dVAE, Ramesh et al., 2021) to produce the labels in the self-supervision objective. Next, the model was fine-tuned for the task of isolated sign language recognition on the gloss-based version (Dafnis et al., 2022) of the WLASL2000 dataset (Li et al., 2020).

The input data was segmented into non-overlapping, padded chunks of 16 frames in order to meet the input requirements of VideoSwin. The outputs were concatenated together.

Following the findings of Uthus et al. (2023) that English pre-trained T5 and fine-tuned for ASL to English translation produces state-of-the-art results, the TTIC team used a T5 model pre-trained on the German Colossal Cleaned Common Crawl (GC4) corpus.[11] They used pre-trained checkpoints from HuggingFace (Wolf et al., 2019). To tokenize the target side, SentencePiece (Kudo and Richardson, 2018) trained on the same data was used to produce a vocabulary of 32,128 tokens.

Their system employs a convolutional layer that is trained to project the sequence of visual features into a single vector per time step. The T5 embeddings layer is replaced by this convolutional layer. The cross-entropy loss was used for the BEVT pre-

[11] https://german-nlp-group.github.io/projects/gc4-corpus.html

training, the ISLR fine-tuning, the text-to-text pre-training as well as for the translation. At inference time, the diverse beam search algorithm (Vijayakumar et al., 2016) with 5 beams, 5 beam groups and a diversity penalty of 1 was used. In contrast to KNOWCOMP, the TTIC team used 8 GPUs to train their system.

6.5 Submission by CASIA

Finally, we received several submissions from the National Laboratory of Pattern Recognition at the Institute of Automation, Chinese Academy of Sciences (submission ID: CASIA). No system paper was submitted and the authors did not provide further information.

7 Evaluation Protocols

We performed both a human (§7.1) and an automatic (§7.2) evaluation of translation quality. Our final system ranking is based on the human evaluation only.

7.1 Human evaluation

Our human evaluation follows the setting we established last year for SLT human evaluation with custom guidelines (Müller et al., 2022), which was originally adapted from the evaluation protocol used at the recent WMT conferences (Kocmi et al., 2022).

Scoring method We employed the source-based direct assessment (DA; Graham et al., 2013; Cettolo et al., 2017) methodology with document context, extended with Scalar Quality Metric (SQM; Freitag et al., 2021). Assessments were performed on a continuous scale between 0 and 100 as in traditional DA but with 0-6 markings on the analogue slider and custom annotator guidelines specifically designed for our task.

As a result of the human evaluation, the systems are ranked from best to worst, after averaging the segment-level DA scores given by the human annotators. In contrast to previous evaluation campaigns (Akhbardeh et al., 2021) which calculate the rankings based on standardized scores (z-scores), we decided to not do so, because the large number of zero-scored items led to a rather skewed standardization scale which affected the calculation of the clusters. We did not make any distinction between segment-level and document-level scores, simply including the latter as additional data for computing the average scores.

After ranking the systems based on their average scores, they are grouped into significance clusters, following the Wilcoxon rank-sum test. Rank ranges give an indication of the translation quality of a system within a cluster and are based on the same head-to-head statistical significance tests.

Inter- and intra-annotator agreement was measured with Fleiss κ (Fleiss, 1971). This should be considered an approximation, noting the concerns of Ma et al. (2017) that kappa coefficients are not suitable for continuous scales. In order to calculate the coefficient, the values have been discretized in seven bins in the scale 0-6, since those were the scores marked on the continuous evaluation bar that was given to the annotators.

Settings of evaluation campaign We used the Appraise evaluation framework[12] (Federmann, 2018) for collecting segment-level judgments. As there were submissions in the DSGS-to-German direction only (§6), we only set up a sign-to-text human evaluation campaign. Annotators were presented with video fragments as source context and translation outputs of a random document fragment from an MT system. The reference translation and the official baseline were included as additional system outputs. Document fragments were created from (up to) twelve consecutive segments. The SRF23 part of the test set was evaluated within the document context. Because the Signsuisse part is a collection of utterances without document boundaries, we presented up to twelve random segments at once but emphasized in the guidelines that those are unrelated and should be assessed independently.

A screenshot of an example annotation in Appraise is presented in Figure 3. The full instructions to evaluators in English and German are listed in Appendix B.

Data and scripts used for generating tasks and computing the final system rankings are publicly available in a Github repository.[13]

We hired three evaluators who are native German speakers and trained DSGS interpreters. All of them had prior experience with evaluation of MT output. Each evaluator was assigned an identical set of annotation tasks comprising the entire test set and all participating systems, including the baseline system and the reference translation. As last year, we did not include any quality control items in the annotation tasks as we had multiple independent

[12] https://github.com/AppraiseDev/Appraise
[13] https://github.com/WMT-SLT/wmt-slt23

Figure 3: A screenshot of an example sign-to-text annotation task in Appraise featuring document-level source-based direct assessment (DA) with scalar quality metrics (SQM) and custom annotator guidelines in German. Taken from Müller et al. (2022).

annotations of the entire test set and because of the very low quality of translations, which would make them indistinguishable from segments with randomly replaced words or phrases used as quality control items.

Feedback from evaluators After completing the evaluation all three evaluators filled out the feedback form we used last year regarding the evaluation procedure and the Appraise platform, where they gave us additional informal feedback.

7.2 Automatic evaluation

As in the previous edition, to complement our human evaluation (which provides the main ranking) we also provide an automatic evaluation. We evaluate the submissions from DSGS into German using three automatic metrics: BLEU (Papineni et al., 2002), chrF (Popović, 2015) and BLEURT (Sellam et al., 2020). We note that learned, semantic metrics correlate better with human judgement (Kocmi et al., 2021), but if they consider the source text as an input (e.g. COMET; Rei et al., 2020), they cannot be used in our context because our source is video and not text. There is no known learned metric which supports sign language videos. We use sacreBLEU (Post, 2018) for BLEU[14] and chrF[15] and the Python library for BLEURT.[16] In all cases, we estimate 95% confidence intervals via bootstrap resampling (Koehn, 2004) with 1000 samples.

8 Results

8.1 Human evaluation

Assessment scores All three evaluators completed all tasks, which gave us three independent judgements for each segment from the official test set. In total, for the output of five systems, we collected 7,800 segment-level and 792 document-level assessment scores, which averages to 1,718 scores per system.

System ranking The official system ranking is presented in Table 5. The significance clusters are indicated with horizontal lines. According to our human evaluation (Table 5), the submission by TTIC has achieved an average score of 0.7 on the scale of 0 to 100, compared to a score of 83.8 for

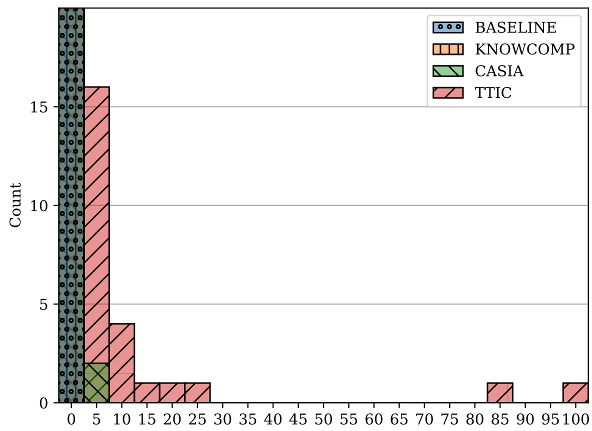

Figure 4: Histogram with the distribution of the system outputs at the DA score scale (x axis) with overlapping semi-transparent bars, discretized into 20 bins. For every segment we include only the average of all ratings. Bin 0, where most ratings belong (up to 496), is cropped to 20 to make the histogram visible.

human translations. The score of TTIC is significantly better than the other systems in the table. All other systems ended up in the same cluster with overall lower translation quality.

Distribution of scores In order to make the distribution of DA scores more interpretable, it is visualized in Figure 4. TTIC had one segment with a score of 99 out of 100, one with 83, one for each of the scores 22, 18 and 15, then 4 segments with a score of about 10, and 16 segments with a score of about 5. CASIA had two segments with a score of about 5. The rest of the segments, including all the outputs from the KNOWCOMP and BASELINE systems, have been given a score very close to 0.

Some example outputs of the highest-scoring translations are listed in Table 6. One can see that TTIC came close to correctly translating the general introductory greetings of the news, but for the rest of the MT ouputs, rated less than 20 out of 100, only a few words match the reference.

Annotator agreement In Table 7 we are reporting intra-annotator agreement for every annotator, measured with Fleiss κ (Fleiss, 1971) over 134 segments which were evaluated twice. (Landis and Koch, 1977; Agresti, 1996). The inter-annotator agreement is $\kappa = 0.80 \pm 0.01$. One can observe that the intra-annotator agreement and all 3 intra-annotator agreements are substantial ($0.61 < \kappa \leq 0.80$) based on Landis and Koch, 1977).

[14] `BLEU|nrefs:1|bs:1000|seed:12345|case:mixed|eff:no|tok:13a|smooth:exp|version: 2.2.0`

[15] `chrF2|nrefs:1|bs:1000|seed:12345|case:mixed|eff:yes|nc:6|nw:0|space:no|version: 2.2.0`

[16] BLEURT v0.0.2 using checkpoint BLEURT-20.

both domains			SRF			Signsuisse		
Rank	Ave.	System	Rank	Ave.	System	Rank	Ave.	System
1	83.829	HUMAN	1	68.809	HUMAN	1	98.630	HUMAN
2	0.669	TTIC	2	1.192	TTIC	2	0.154	TTIC
3-5	0.024	CASIA	3-4	0.046	CASIA	3-5	0.008	BASELINE
3-5	0.008	BASELINE	3-5	0.009	BASELINE	3-5	0.007	KNOWCOMP
3-5	0.005	KNOWCOMP	4-5	0.002	KNOWCOMP	3-5	0.003	CASIA

Table 5: Official results of the WMT23 Sign Language Translation task for translation from Swiss German Sign Language to German. Systems are ordered by averaged (non-standardized) human score in the percentage scale. Lines indicate clusters according to a Wilcoxon rank-sum test $p < 0.05$.

score	system	testset	doc	seg		text
99.3	TTIC	SRF	0	0	hyp:	Guten Abend, meine Damen und Herren, willkommen zur "Tagesschau".
					ref:	Guten Abend, meine Damen und Herren, willkommen zur "Tagesschau".
83.3	TTIC	SRF	0	1	hyp:	Heute mit diesen Themen:
					ref:	Das macht heute Montag Schlagzeilen:
18.7	TTIC	SRF	23	9	hyp:	Der US-Präsident ist heute zu Gast bei "10vor10".
					ref:	Wesentliches gibt es auch heute bei "10vor10".
16.3	TTIC	SRF	18	0	hyp:	Und auch für EU-Bürger, die in die Schweiz einreisen wollen, soll es verschärfte Einreiseregeln geben.
					ref:	Auch die EU will nun ihre Bürger vom Kreuzfahrtschiff zurückholen, denn man misstraut Japans Krisenmanagement.
12.0	TTIC	SRF	14	2	hyp:	Die Leute müssen sich Gedanken machen, wie sie die Zukunft meistern können.
					ref:	Das muss sich ändern, sind sich die EU-Aussenminister einig.
11.0	TTIC	SS	18	5	hyp:	Der Film kann auf YouTube angeschaut werden.
					ref:	Dieser Film ist spannend und interessant.
8.3	TTIC	SRF	15	4	hyp:	Tausende Menschen sind seither ohne Hilfe von aussen ausgewandert.
					ref:	Über 70'000 Menschen haben sich bis heute mit dem neuen Coronavirus infiziert.
5.0	CASIA	SRF	1	1	hyp:	Die Temperaturen steigen in der Schweiz.
					ref:	Und morgen gibt es sonnige Phasen bei Temperaturen um 9 °C.

Table 6: Examples of some of the highest-scoring translations in the test set. hyp=MT outputs, ref=human translation

annotator	kappa
A	0.80±0.05
B	0.80±0.06
C	0.79±0.06

Table 7: Intra-annotator agreement based on the Fleiss κ coefficient for reliability of agreement (with scores discretized in the scale 0-6).

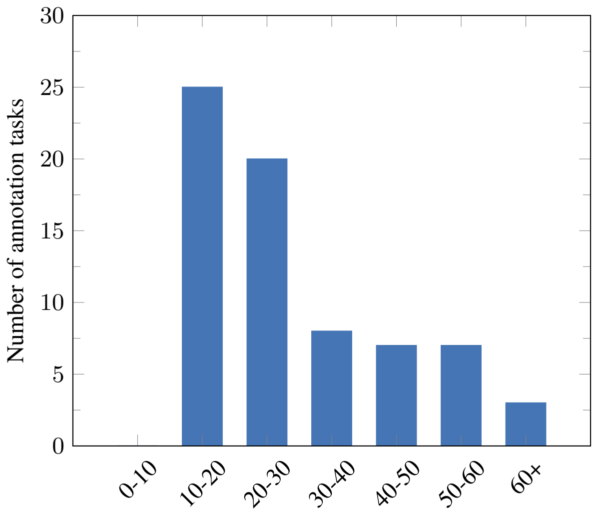

Figure 5: Number of task completion times (a task consists of 100 segments) grouped into 10-minute buckets, after removing top and bottom 5-percentiles.

Evaluation speed A single task requiring providing 100 segment-level and about 12 document-level scores took on average 29 minutes to complete, after excluding 5% of slowest and fastest task annotations. The majority of tasks were finished in between 10 and 30 minutes as shown in Figure 5. This is substantially faster than last year, which averaged around 45 minutes per task.

Feedback from evaluators After completing the evaluation all three evaluators filled in a form meant for feedback regarding the evaluation procedure and the Appraise platform. All evaluators gave us additional informal feedback.

In general, evaluators reported that their experience with Appraise was positive (two of them had used Appraise before), and that our instructions were clear. All of them would be willing to do similar work in the future. They found source videos understandable and the documents or segments given were neither too long nor too short. The general method of assessing translations (DA with SQM) was not found difficult nor stressful, but on the contrary annotators thought it was efficient, simple, fast and practical.

Concerning Appraise development, nobody experienced technical problems, which is an improvement over last year, when two people experienced major technical issues. Evaluators suggested that the user interface could be improved in some places. For instance, automatically playing videos could make evaluations more efficient, the videos should be bigger by default, there should be more keyboard shortcuts and there should be a quick way to give a low score to an entire document.

As explained in more detail below (§9.3), and similar to last year, evaluators told us that some videos do not have ideal cuts, in the sense that the beginning or end are slightly cut off. This is perhaps inevitable in continuous signing, or a problem in our manual alignment process.

Full responses to the feedback form submitted by evaluators are listed in Appendix C.

8.2 Automatic evaluation

Table 8 summarises the results of the automatic evaluation. In general, the translation of the Signsuisse subset (SS) and the SRF23 subset seem to have a similar complexity, especially according to chrF and BLEURT evaluation scores. BLEU, on the other hand, shows higher translation quality for SRF in selected systems by CASIA and TTIC. Both teams are able to significantly outperform the baseline system according to the three evaluation metrics. TTIC achieves the best scores with their primary submission TTIC.423. Although chrF points out another of their submissions as the best system, the difference with respect to the primary submission is not statistically significant.

9 Discussion

9.1 General translation quality

Overall, all systems perform poorly in our shared task, as there is an extreme difference in average score between all systems and the human reference translation. The systems exhibit well-known problems of natural language generation such as overfitting to few high-probability hypotheses and hallucination (Lee et al., 2018; Raunak et al., 2021).

The best submitted system in the best case achieves an average score of about 1 out of 100 (where the human translation achieved 69 out of 100), which indicates that current automatic translations are not usable in practice, unlike spoken language MT where in specific scenarios experiments have shown systems to be on par with human

	BLEU			chrF			BLEURT		
Submission	all	SS	SRF23	all	SS	SRF23	all	SS	SRF23
BASELINE	0.09±0.03	0.15±0.06	0.10±0.05	12.4±0.4	12.2±0.5	12.5±0.5	0.072±0.003	0.083±0.005	0.060±0.005
CASIA.426	0.38±0.20	0.16±0.04	0.52±0.28	14.6±0.4	14.2±0.5	14.8±0.7	0.148±0.006	0.143±0.008	0.152±0.007
CASIA.427	0.39±0.20	0.13±0.05	0.52±0.28	14.2±0.5	13.4±0.5	14.8±0.7	0.162±0.006	0.171±0.009	0.152±0.007
CASIA.428	0.16±0.07	0.16±0.04	0.20±0.10	13.5±0.4	14.2±0.5	13.0±0.5	0.156±0.005	0.143±0.008	0.168±0.007
CASIA.429	0.38±0.20	0.15±0.06	0.52±0.28	14.3±0.4	13.5±0.5	14.8±0.7	0.175±0.006	0.197±0.008	0.152±0.007
CASIA.430	0.33±0.16	0.15±0.10	0.52±0.28	14.7±0.4	14.6±0.5	14.8±0.7	0.166±0.006	0.179±0.008	0.152±0.007
CASIA.431	0.13±0.06	0.15±0.10	0.14±0.03	14.5±0.4	14.6±0.5	14.4±0.6	0.169±0.006	0.179±0.008	0.159±0.008
CASIA.432	0.37±0.19	0.11±0.05	0.52±0.28	14.4±0.4	13.7±0.5	14.8±0.7	0.172±0.006	0.190±0.008	0.152±0.007
KNOWCOMP.418	0.06±0.03	0.07±0.03	0.09±0.04	6.2±0.3	6.9±0.5	5.7±0.5	0.077±0.005	0.080±0.007	0.073±0.007
KNOWCOMP.419	0.07±0.05	0.06±0.02	0.11±0.09	7.6±0.3	8.2±0.4	7.2±0.4	0.083±0.005	0.084±0.007	0.081±0.007
TTIC.417	0.56±0.46	0.30±0.14	0.29±0.13	15.9±0.5	16.6±0.8	15.3±0.6	0.222±0.010	0.231±0.011	0.210±0.015
TTIC.420	0.78±0.83	0.21±0.04	0.17±0.02	16.0±0.5	16.2±0.6	15.5±0.6	0.224±0.010	0.228±0.011	0.216±0.015
TTIC.421	0.21±0.09	0.13±0.06	0.29±0.13	13.2±0.4	13.3±0.5	13.2±0.6	0.087±0.006	0.078±0.006	0.095±0.010
TTIC.422	0.77±0.74	0.22±0.13	0.29±0.12	17.3±0.5	16.7±0.6	17.4±0.6	0.239±0.010	0.230±0.011	0.245±0.015
TTIC.423	1.03±0.87	0.21±0.03	0.69±0.46	17.0±0.6	16.2±0.7	17.2±0.7	0.243±0.010	0.236±0.011	0.246±0.013
TTIC.424	0.79±0.74	0.24±0.12	0.33±0.14	17.2±0.5	16.6±0.7	17.5±0.7	0.236±0.009	0.228±0.011	0.241±0.015
TTIC.425	0.74±0.79	0.14±0.06	0.23±0.10	16.3±0.6	16.0±0.7	16.3±0.7	0.205±0.009	0.194±0.010	0.214±0.014

Table 8: Automatic evaluation of all the submission for the full WMT-SLT test set (all), the Signsuisse subset (SS) and the SRF23 subset. Mean and 95% confidence intervals obtained via bootstrap resampling are shown. Primary submissions manually evaluated are boldfaced.

translation (Hassan et al., 2018; Popel et al., 2020). This assessment of general translation quality is unchanged from last year, see Müller et al. (2022) for potential reasons that still apply to the current shared task.

9.2 No submissions for spoken-to-signed translation directions

No teams participated in a track where a sign language is the target language (§3). We believe this could be due to the fact that generating sign language may appear considerably harder to participants. The problem of signed-to-spoken translation fits well into existing translation paradigms and toolkits, because using arbitrary features on the source side is easier than generating arbitrary numerical data (such as a video). Decoding text on the target side is considerably easier and more well-defined in NLP than decoding a video or similar data structure.

We thought that providing a baseline system for spoken-to-signed translation (§6.2) may help lower the barriers to entry but clearly, more measures are needed. A different hypothesis is that our shared task in its current form does not appeal to scientists working in the field of sign language generation or avatar technology. They may have felt alienated by aspects of the shared task which are familiar to MT researchers, but would need more explanation or introduction for people from neighboring fields.

9.3 Low scores of human translations

When looking at the domain-specific results (Table 5b and c), we observe that the human translation in SRF was ranked considerably lower than Signsuisse (69% against 98%). This difference warrants further investigation, as does the fact that a percentage of 69% is by itself rather low. We explain potential reasons for this below, attributing the difference to the way the corpora were generated.

Interpretation vs. translation SRF is partially generated as live interpretation of the spoken TV shows (spoken-to-sign), where interpreters are under time pressure. Due to specific efficiency strategies they occasionally omit content to keep up with the spoken audio. Therefore, since here we are evaluating the performance of the systems in the opposite direction (sign-to-spoken) it may as well very often be that the content of the interpretation does not match the one of the written or spoken sentence. However, as explained in Section 4, the Signsuisse part of the testset derives from a lexicon, containing sentences recorded as examples of particular lexicon entries. Since these have been generated for the purpose of being included in the lexicon, the accuracy of the translation is expected to be much higher than the one achieved within live interpretation.

Video editing issues The measured bad human performance on SRF may also be explained by the fact that the video cuts are sometimes not ideal, i.e. the beginning or end of an SL utterance is cut off, as noted by our evaluators. This may have occurred because segmenting continuous signing is difficult and there is no ideal way to separate seamless transitions.

In the future these problems could perhaps be mitigated by including more frames from the left and right border of a video clip, or simply discarding sentences with unclear boundaries.

Role of discourse context A third reason may be that SLs are probably more dependent on context than spoken languages, e.g. because of index signs. This means that evaluating an isolated SL utterance (the equivalent of one sentence in a spoken language) may lead to low scores. This is a phenomenon that would more likely occur in a news report of SRF, as compared to the isolated example sentences of Signsuisse.

Contrary to what was observed for the evaluation of the human translation, the two submitted MT systems TTIC and CASIA perform significantly better on SRF than on Signsuisse. Here we may provide the assumption, that since the amount of training sentences from SRF is bigger than the ones from Signsuisse, the systems are optimized better for that domain. Additionally, it has been noted that in interpretation settings similar to the ones of SRF, the linguistic characteristics of the signing may be more closely related to German than in an offline translation setting, such as the one in Signsuisse.

9.4 Quality of training data and unexplored potential

Compared to last year we offered considerably more training data (hundreds of hours worth of video compared to dozens last year; §4.2). However, while last year all training data was manually corrected, this year we offered the data as-is. The SRF23 training data is best understood as a comparable corpus, or web-crawled parallel corpus including various types of noise (Khayrallah and Koehn, 2018). For instance, the time stamps of the German subtitles are more aligned with the audio signal present in the broadcast and do not account for the delay of live-interpreted signing. Any naive extraction of parallel examples from SRF23 without any alignment tools or shifting subtitle times will result in noisy training data.

As far as we know no participant investigated ways to improve the alignments automatically, which is perhaps because we did not explain this well in our online documentation. One reason for this may be that we did not make it clear enough to participants that one of our training corpora is effectively un-aligned. But essentially, it means there is unexplored potential in improving or filtering the training data instead of training on the raw corpora.

9.5 Limitations of shared task setup

The limitations we identified in last year's findings paper still apply. Briefly, the limitations concern the lack of generalization across signers, the favourable recording conditions of our sign language data and interpretation vs. translation setups. See Müller et al. (2022) for a more comprehensive description.

10 Conclusion and future directions

In this paper we present the second WMT Shared Task on Sign Language Translation (WMT-SLT23). We consider automatic sign language translation, and sign language processing in general, to be of wide public interest and to have a high potential impact in a societal and academic sense (§2).

Compared to last year we ran our shared task for three language pairs instead of one, we distributed considerably more training data (albeit with a higher amount of noise) and we put more emphasis on scenarios where sign languages are the target language.

Four teams participated in the second edition of the shared task. Overall, we observed low system performance with an average human evaluation score of about 1 out of 100 (for the best-performing system), which is not usable in practice. The main reasons for this outcome are a lack of usable training data, a modality gap (considering that most existing work in MT is based on text) and a lack of basic NLP tools specifically for sign languages.

Future of the shared task After two successful iterations the shared task is now well established, in the sense that suitable protocols are in place for human and automatic evaluation, reasonable baseline systems exist, as well as several training corpora and official WMT test sets.

So far our shared tasks have certainly helped to paint a more realistic picture of the translation quality of state-of-the-art systems, but they have not led to any major technical innovation. This may be because technologies more fundamental than machine translation do not exist for sign languages, or are not reliable enough. For this reason we will consider running shared tasks on more fundamental problems in SLP such as alignment, segmentation, or automatic filtering of parallel corpora.

In the future we could also try to shift the focus away from interpreted news broadcast material as the basis for training and test data. A major challenge to overcome is that interpreted material is available in larger amounts, while signing produced by conventional, off-line translation or produced by native signers is harder to come by. Nevertheless, using non-interpreted material largely avoids alignment shifts in the training data and leads to higher scores for the human translations of the test data, among other advantages.

11 Ethical statement

Within this shared task, two main ethical considerations emerge: the potential impact of SL technology on target users and privacy considerations.

Research in sign language processing, if not executed carefully, may inadvertently cause harm to end users, especially members of deaf communities. Hearing scientists should refrain from prescribing what sort of language technology should be accepted by deaf or hard-of-hearing individuals and should avoid claiming that their approach "solves" any particular problem. Ideally, research of this nature should include deaf and hard-of-hearing people, not only at evaluation time but in the entire development cycle (Fox et al., 2023).

Secondly, there is a concern for the privacy of individuals depicted in SLP datasets. For the specific use case of sign language data, proper anonymisation is impossible, since identifying details such as facial expressions are crucial for sign language communication. We have obtained written permission of all individuals shown in our datasets. Storing and processing pose estimation features instead of raw videos may be an alternative that provides anonymity (and has other generalization effects such as ignoring differences in race, gender, clothing, background, etc.). However, in our shared task and related literature, (Moryossef et al., 2021; Tarrés et al., 2023) video features outperform pose features.

Acknowledgements

The organizing committee acknowledges funding from the following projects: the EU Horizon 2020 projects EASIER (grant agreement number 101016982) and SignON (101017255), the Swiss Innovation Agency (Innosuisse) flagship IICT (PFFS-21-47) and the German Ministry of Education and Research (BMBF) through the project SocialWear (01IW20002).

Finally, we would like to extend heartfelt thanks to the DSGS interpreters who performed our human evaluation: Heidi Stocker, Janine Criblez and Tanja Joseph.

References

Alan Agresti. 1996. *An introduction to categorical data analysis*, volume 135. Wiley New York.

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 Conference on Machine Translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Samuel Albanie, Gül Varol, Liliane Momeni, Triantafyllos Afouras, Joon Son Chung, Neil Fox, and Andrew Zisserman. 2020. Bsl-1k: Scaling up co-articulated sign language recognition using mouthing cues. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 35–53. Springer.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-Scale Acquisition of Parallel Corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Elise Bertin-Lemée, Annelies Braffort, Camille Challant, Claire Danet, and Michael Filhol. 2023. Example-based machine translation from textto a hierarchical representation of sign language. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 21–30, Tampere, Finland. European Association for Machine Translation.

Claudia Bianchini and Fabrizio Borgia. 2012. Writing sign languages: analysis of the evolution of the sign-writing system from 1995 to 2010, and proposals for future developments. In *Proceedings of the Intl Jubilee Congress of the Technical University of Varna*, pages 118–123.

Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, Christian Vogler, and Meredith Ringel Morris. 2019. Sign Language Recognition, Generation, and Translation: An Interdisciplinary Perspective. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '19, pages 16–31, New York, NY, USA. Association for Computing Machinery.

Hannah Bull, Annelies Braffort, and Michèle Gouiffès. 2020a. Mediapi-skel-a 2d-skeleton video database of french sign language with aligned french subtitles. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 6063–6068.

Hannah Bull, Michèle Gouiffès, and Annelies Braffort. 2020b. Automatic segmentation of sign language into subtitle-units. In *European Conference on Computer Vision*, pages 186–198. Springer.

Necati Cihan Camgöz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural Sign Language Translation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7784–7793.

Necati Cihan Camgöz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020a. Multi-channel transformers for multi-articulatory sign language translation. In *European Conference on Computer Vision*, pages 301–319.

Necati Cihan Camgöz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020b. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10023–10033.

Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. 2019. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Niehues Jan, Stüker Sebastian, Sudoh Katsuitho, Yoshino Koichiro, and Federmann Christian. 2017. Overview of the IWSLT 2017 evaluation campaign. In *International Workshop on Spoken Language Translation*, pages 2–14.

Onno Crasborn. 2006. Nonmanual structures in sign language. *Encyclopedia of Language and Linguistics*, 8:668–672.

Konstantinos M Dafnis, Evgenia Chroni, Carol Neidle, and Dimitri Metaxas. 2022. Bidirectional skeleton-based isolated sign recognition using graph convolutional networks. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7328–7338.

Daria Dayter. 2019. Collocations in non-interpreted and simultaneously interpreted english: a corpus study. In *New empirical perspectives on translation and interpreting*, pages 67–91. Routledge.

Mathieu De Coster, Dimitar Shterionov, Mieke Van Herreweghe, and Joni Dambre. 2022. Machine translation from signed to spoken languages: State of the art and challenges. *arXiv preprint arXiv:2202.03086*.

Mirella De Sisto, Dimitar Shterionov, Irene Murtagh, Myriam Vermeerbergen, and Lorraine Leeson. 2021. Defining Meaningful Units. Challenges in Sign Segmentation and Segment-Meaning Mapping (short paper). In *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*, pages 98–103, Virtual. Association for Machine Translation in the Americas.

Mirella De Sisto, Vincent Vandeghinste, Santiago Egea Gómez, Mathieu De Coster, Dimitar Shterionov, and Horacio Saggion. 2022. Challenges with sign language datasets for sign language recognition and translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2478–2487, Marseille, France. European Language Resources Association.

Elisabeth Engberg-Pedersen. 1993. *Space in Danish Sign Language: The Semantics and Morphosyntax of the Use of Space in a Visual Language*. SIGNUM-Press.

Christian Federmann. 2018. Appraise Evaluation Framework for Machine Translation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 86–88, Santa Fe, New Mexico. Association for Computational Linguistics.

Michael Filhol. 2020. Elicitation and Corpus of Spontaneous Sign Language Discourse Representation Diagrams. In *Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages at Language Resources and Evaluation Conference*, pages 53–60. European Language Resources Association (ELRA).

Joseph L. Fleiss. 1971. Measuring Nominal Scale Agreement Among Many Raters. *Psychological bulletin*, 76(5):378.

Jens Forster, Christoph Schmidt, Oscar Koller, Martin Bellgardt, and Hermann Ney. 2014. Extensions of the Sign Language Recognition and Translation Corpus RWTH-PHOENIX-Weather. In *Proceedings of the Ninth International Conference on Language*

Resources and Evaluation (LREC'14), pages 1911–1916, Reykjavik, Iceland. European Language Resources Association (ELRA).

Neil Fox, Bencie Woll, and Kearsy Cormier. 2023. Best practices for sign language technology research. *Universal Access in the Information Society*, pages 1–9.

Markus Freitag, 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. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous Measurement Scales in Human Evaluation of Machine Translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Thomas Hanke. 2004. Hamnosys - representing sign language data in language resources and language processing contexts. In *LREC 2004, Workshop proceedings : Representation and processing of sign languages*, pages 1–6. Paris : ELRA.

Thomas Hanke, 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.

Thomas Hanke and Jakob Storz. 2008. ilex–a database tool for integrating sign language corpus linguistics and sign language lexicography. In *sign-lang@ LREC 2008*, pages 64–67. European Language Resources Association (ELRA).

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, et al. 2018. Achieving human parity on automatic chinese to english news translation. *arXiv preprint arXiv:1803.05567*.

Mert Inan, Yang Zhong, Sabit Hassan, Lorna Quandt, and Malihe Alikhani. 2022. Modeling intensification for sign language generation: A computational approach. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2897–2911, Dublin, Ireland. Association for Computational Linguistics.

Zifan Jiang, Amit Moryossef, Mathias Müller, and Sarah Ebling. 2023a. Signsuisse dsgs/lsf/lis lexicon.

Zifan Jiang, Amit Moryossef, Mathias Müller, and Sarah Ebling. 2023b. Srf dsgs daily news broadcast: video and original subtitle data.

Trevor Johnston. 2010. From archive to corpus: Transcription and annotation in the creation of signed language corpora. *International Journal of Corpus Linguistics*, 15(1):106–131.

Trevor Johnston. 2011. Lexical Frequency in Sign Languages. *The Journal of Deaf Studies and Deaf Education*, 17(2):163–193.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Huda Khayrallah and Philipp Koehn. 2018. On the impact of various types of noise on neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 74–83, Melbourne, Australia. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Sang-Ki Ko, Chang Jo Kim, Hyedong Jung, and Choongsang Cho. 2019. Neural sign language translation based on human keypoint estimation. *Applied Sciences*, 9(13):2683.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To Ship or Not to Ship: An Extensive Evaluation of Automatic Metrics for Machine Translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical Significance Tests for Machine Translation Evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Oscar Koller. 2020. Quantitative survey of the state of the art in sign language recognition. *arXiv preprint arXiv:2008.09918*.

Maria Kopf, Marc Schulder, and Thomas Hanke. 2021. Overview of Datasets for the Sign Languages of Europe.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

J R Landis and G G Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174.

Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. 2018. Hallucinations in Neural Machine Translation.

Lorraine Leeson. 2005. Making the effort in simultaneous interpreting. *Topics in Signed Language Interpreting: Theory and Practice, ed. by Terry Janzen.–Amsterdam*.

Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. 2020. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1459–1469.

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. 2022. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211.

Camillo Lugaresi, 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. *CoRR*, abs/1906.08172.

Qingsong Ma, Yvette Graham, Timothy Baldwin, and Qun Liu. 2017. Further investigation into reference bias in monolingual evaluation of machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2466–2475, Copenhagen, Denmark. Association for Computational Linguistics.

Caio DD Monteiro, Christy Maria Mathew, Ricardo Gutierrez-Osuna, and Frank Shipman. 2016. Detecting and identifying sign languages through visual features. In *2016 IEEE International Symposium on Multimedia (ISM)*, pages 287–290. IEEE.

Sara Morrissey. 2011. Assessing three representation methods for sign language machine translation and evaluation. In *Proceedings of the 15th annual meeting of the European Association for Machine Translation (EAMT 2011), Leuven, Belgium*, pages 137–144.

Amit Moryossef and Yoav Goldberg. 2021. Sign Language Processing. `https://sign-language-processing.github.io/`.

Amit Moryossef and Mathias Müller. 2021. pose-format: Library for viewing, augmenting, and handling .pose files. `https://github.com/AmitMY/pose-format`.

Amit Moryossef, 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`.

Amit Moryossef, Ioannis Tsochantaridis, Joe Dinn, Necati Cihan Camgöz, Richard Bowden, Tao Jiang, Annette Rios, Mathias Müller, and Sarah Ebling. 2021. Evaluating the immediate applicability of pose estimation for sign language recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 10166v1. 2021 ChaLearn Looking at People Sign Language Recognition in the Wild Workshop at CVPR.

Mathias Müller, 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)*, pages 744–772, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Mathias Müller, 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)*, pages 682–693, Toronto, Canada. Association for Computational Linguistics.

Ellen Ormel and Onno Crasborn. 2012. Prosodic correlates of sentences in signed languages: A literature review and suggestions for new types of studies. *Sign Language Studies*, 12(2):279–315.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Kishore Papineni, 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*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Pamela Perniss, Asli Özyürek, and Gary Morgan. 2015. The influence of the visual modality on language structure and conventionalization: Insights from sign language and gesture. *Topics in Cognitive Science*, 7(1):2–11.

Elena Pizzuto and Paola Pietrandrea. 2001. The Notation of Signed Texts: Open Questions and Indications for Further Research. *Sign Language & Linguistics*, 4:29–45.

Martin Popel, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtskỳ. 2020. Transforming Machine Translation: a Deep Learning System Reaches News Translation Quality Comparable to Human Professionals. *Nature communications*, 11(1):1–15.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The Curious Case of Hallucinations in Neural Machine Translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1172–1183, Online. Association for Computational Linguistics.

Ricardo Rei, 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)*, pages 2685–2702, Online. Association for Computational Linguistics.

Marcelo Sandoval-Castaneda, Yanhong Li, Bowen Shi, Diane Brentari, Karen Livescu, and Gregory Shakhnarovich. 2023. TTIC's Submission to WMT-SLT 23. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Ben Saunders, Necati Cihan Camgöz, and Richard Bowden. 2020a. Adversarial training for multi-channel sign language production. In *The 31st British Machine Vision Virtual Conference*. British Machine Vision Association.

Ben Saunders, Necati Cihan Camgöz, and Richard Bowden. 2020b. Everybody sign now: Translating spoken language to photo realistic sign language video. *arXiv preprint arXiv:2011.09846*.

Ben Saunders, Necati Cihan Camgöz, and Richard Bowden. 2020c. Progressive transformers for end-to-end sign language production. In *European Conference on Computer Vision*, pages 687–705.

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2022. Signing at Scale: Learning to Co-Articulate Signs for Large-Scale Photo-Realistic Sign Language Production. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning Robust Metrics for Text Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Rico Sennrich and Biao Zhang. 2019. Revisiting Low-Resource Neural Machine Translation: A Case Study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 211–221, Florence, Italy. Association for Computational Linguistics.

Bowen Shi, Diane Brentari, Gregory Shakhnarovich, and Karen Livescu. 2022. Open-domain sign language translation learned from online video. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6365–6379, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Anita Slonimska, Asli Özyürek, and Olga Capirci. 2021. Using Depiction for Efficient Communication in LIS (Italian Sign Language). *Language and Cognition*, 13(3):367–396.

Stephanie Stoll, Necati Cihan Camgöz, Simon Hadfield, and Richard Bowden. 2020. Text2Sign: Towards Sign Language Production Using Neural Machine Translation and Generative Adversarial Networks. *International Journal of Computer Vision*, 128(4):891–908.

Valerie Sutton. 1990. *Lessons in sign writing*. SignWriting.

Laia Tarrés, 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 (CVPR) :Workshops*.

David Uthus, Garrett Tanzer, and Manfred Georg. 2023. Youtube-asl: A large-scale, open-domain american sign language-english parallel corpus.

Gül Varol, Liliane Momeni, Samuel Albanie, Triantafyllos Afouras, and Andrew Zisserman. 2021. Read and attend: Temporal localisation in sign language videos. In *CVPR*.

Gül Varol, Liliane Momeni, Samuel Albanie, Triantafyllos Afouras, and Andrew Zisserman. 2022. Scaling up sign spotting through sign language dictionaries. *International Journal of Computer Vision*, 130(6):1416–1439.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Carla Viegas, Mert Inan, Lorna Quandt, and Malihe Alikhani. 2023. Including facial expressions in contextual embeddings for sign language generation. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 1–10, Toronto, Canada. Association for Computational Linguistics.

Ashwin K. Vijayakumar, Michael Cogswell, Ramprasaath R. Selvaraju, Qing Sun, Stefan Lee, David J. Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *CoRR*, abs/1610.02424.

Harry Walsh, Ben Saunders, and Richard Bowden. 2022. Changing the representation: Examining language representation for neural sign language production. In *LREC 2022*.

Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li, and Chunhua Shen. 2020. SOLOv2: Dynamic and Fast Instance Segmentation. In *Advances in Neural Information Processing Systems*, volume 33, pages 17721–17732. Curran Associates, Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Rosalee Wolfe, John C. McDonald, Thomas Hanke, Sarah Ebling, Davy Van Landuyt, Frankie Picron, Verena Krausneker, Eleni Efthimiou, Evita Fotinea, and Annelies Braffort. 2022. Sign language avatars: A question of representation. *Information*, 13(4):206.

Bencie Woll. 2013. 9091 The History of Sign Language Linguistics. In *The Oxford Handbook of the History of Linguistics*. Oxford University Press.

Baixuan Xu, Haochen Shi, Tianshi Zheng, Qing Zong, Weiqi Wang, Zhaowei Wang, and Yangqiu Song. 2023. KnowComp Submission for WMT23 Sign Language Translation Task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. 2021. Including Signed Languages in Natural Language Processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7347–7360, Online. Association for Computational Linguistics.

Kayo Yin and Jesse Read. 2020. Better sign language translation with stmc-transformer. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5975–5989.

A Details on shared task data and submission

A.1 Data resources

Direct download links: https://www.swissubase.ch/en/catalogue/studies/20452/19173/datasets/2327/2705/overview
Signsuisse lexicon (release 2.0): https://www.swissubase.ch/en/catalogue/studies/20452/19280/datasets/2350/2715/overview
SRF corpus poses and segmented subtitles (release 1.0): https://www.swissubase.ch/en/catalogue/studies/20452/19280/datasets/2343/2721/overview
Test sources as a tar ball (release 2.0): https://files.ifi.uzh.ch/cl/archiv/2023/easier/wmtslt/test_sources.v2.0.tar.gz
Test sources in WMT XML format for submissions: https://files.ifi.uzh.ch/cl/archiv/2023/easier/wmtslt/xml/

A.2 XML submission schema

```xml
<?xml version='1.0' encoding='utf-8'?>
<dataset id="slttest2022.de-dsgs">
  <doc origlang="de" id="srf.0">
    <src lang="de">
      <p>
        <seg id="0">Guten Abend meine Damen und Herren - willkommen zur "Tagesschau".</seg>
      </p>
    </src>
    <hyp system="YOUR SYSTEM NAME" language="dsgs">
      <p>
        <seg id="0"> https://www.your_hosting.com/your_url_for_this_segment </seg>
      </p>
    </hyp>
  </doc>
</dataset>
```

B Appraise instructions to human evaluators

B.1 Sign-to-text direction

B.1.1 English

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 on 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.

Please score the overall document translation quality (you can score the whole document only after scoring all individual sentences first). 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.

B.1.2 German

Unten sehen Sie ein Dokument mit 10 Sätzen in Deutschschweizer Gebärdensprache (DSGS) (linke Spalten) und die entsprechenden möglichen Übersetzungen auf Deutsch (rechte Spalten). Bewerten Sie jede mögliche Übersetzung des Satzes im Kontext des Dokuments. Sie können bereits bewertete Sätze jederzeit durch Anklicken eines Eingabevideos erneut aufrufen und die Bewertung aktualisieren.

Bewerten Sie die Übersetzungsqualität auf einer kontinuierlichen Skala mit Hilfe der nachfolgend beschriebenen Qualitätsstufen:

- 0: Unsinn/Bedeutung nicht erhalten: Fast alle Informationen zwischen Übersetzung und Eingabevideo sind verloren gegangen. Die Grammatik ist irrelevant.
- 2: Ein Teil der Bedeutung ist erhalten: Die Übersetzung behält einen Teil der Bedeutung der Quelle bei, lässt aber wichtige Teile aus. Die Erzählung ist aufgrund von grundlegenden Fehlern schwer zu verstehen. Die Grammatik kann mangelhaft sein.
- 4: Der grösste Teil der Bedeutung ist erhalten und es gibt nur wenige Grammatikfehler: Die Übersetzung behält den grössten Teil der Bedeutung der Quelle bei. Sie kann einige Grammatikfehler oder kleinere kontextuelle Unstimmigkeiten aufweisen.
- 6: Perfekte Bedeutung und Grammatik: Die Bedeutung der Übersetzung stimmt vollständig mit der Quelle und dem umgebenden Kontext (falls zutreffend) überein. Auch die Grammatik ist korrekt.

Bitte bewerten Sie die Übersetzungsqualität des gesamten Dokuments. (Sie können das Dokument erst bewerten, nachdem Sie zuvor alle Sätze einzeln bewertet haben.) Bewerten Sie die Übersetzungsqualität auf einer kontinuierlichen Skala mit Hilfe der nachfolgend beschriebenen Qualitätsstufen:

- 0: Unsinn/Bedeutung nicht erhalten: Fast alle Informationen zwischen Übersetzung und Eingabevideo sind verloren gegangen. Die Grammatik ist irrelevant.
- 2: Ein Teil der Bedeutung ist erhalten: Die Übersetzung behält einen Teil der Bedeutung der Quelle bei, lässt aber wichtige Teile aus. Die Erzählung ist aufgrund von grundlegenden Fehlern schwer zu verstehen. Die Grammatik kann mangelhaft sein.
- 4: Der grösste Teil der Bedeutung ist erhalten und es gibt nur wenige Grammatikfehler: Die Übersetzung behält den grössten Teil der Bedeutung der Quelle bei. Sie kann einige Grammatikfehler oder kleinere kontextuelle Unstimmigkeiten aufweisen.
- 6: Perfekte Bedeutung und Grammatik: Die Bedeutung der Übersetzung stimmt vollständig mit der Quelle und dem umgebenden Kontext (falls zutreffend) überein. Auch die Grammatik ist korrekt.

C Feedback from evaluators

Tables 9 and 10 detail for each evaluator the feedback answers and comments regarding the human evaluation procedure and the Appraise system. All three evaluators submitted a response.

	Answer 1	**Answer 2**	**Answer 3**
What is your experience in assessing machine translation outputs?			
	Low: I have done it once or a long time ago	Moderate: I have done it a few times	Low: I have done it once or twice before, or a long time ago
Please specify how much you agree or disagree with the following statements.			
Generally, my experience with the tool was positive	Agree	Agree	Agree
Instructions were clear	Neutral	Strongly agree	Strongly agree
Quality levels 0-6 were helpful to me	Neutral	Neutral	Agree
Source videos were understandable	Strongly agree	Agree	Strongly Agree
There was too much repetitiveness	Strongly agree	Neutral	Strongly agree
Documents were too long	Disagree	Disagree	Neutral
Segments were too short	Disagree	Disagree	Disagree
In some cases, the context was insufficient	Neutral	Neutral	Disagree
I experienced technical issues	Neutral	Neutral	Disagree
I would be willing to do similar work in the future	Agree	Agree	Agree
This evaluation campaign featured the Direct Assessment with Scalar Quality Metrics method. What do you think about this method? On a scale between -3 (negative) and 3 (positive) it was...			
difficult/easy	+1	+3	+3
stressful/relaxed	0	+3	+2
laborious/effortless	+2	+2	-2
slow/fast	+2	+2	0
inefficient/efficient	+2	+2	+2
boring/exciting	-1	+2	0
complicated/simple	+1	+2	+3
annoying/enjoyable	-1	+2	0
limiting/creative	-1	0	0
impractical/practical	0	+2	+3

Table 9: Feedback from evaluators about the human evaluation setup and the Appraise platform.

Answer 1		Answer 2	Answer 3
Please provide more details related to the statements above that you think can be useful to us. What was most troublesome? What could we improve?			
(original in German) - Ich hätte ein grösseres Video geschätzt (ohne dass ich das jedes Mal aktiv anklicken muss) > Z.B. bei Klicken auf Play, automatische Vergrösserung und bei Ende der Wiedergabe automatisch zurück auf die Skala. - Die Videoschnitte waren - v.a. bei einem Modell (langer Lag!) - sehr schlecht. Video und Text stimmten deshalb oft nicht überein. Schwierig für die Beurteilung! - Es kam oft vor, dass ganze Dokumente schon auf einen Blick als "komplett falsch" ersichtlich waren (Texte komplett unverständlich). Da wäre es hilfreich, wenn man ein gesamtes Dokument als "ROT" beurteilen könnte, ohne jedes einzelne Video zu beurteilen.	(translated into English) - I would have appreciated a larger video (without having to actively click that every time) > E.g. when clicking play, automatic enlargement and at the end of playback automatically back to the scale. - The video cuts were - especially with one model (long lag!) - very bad. Video and text therefore often did not match. Difficult for the evaluation! - It often happened that whole documents appeared at a glance as "completely wrong" (texts completely incomprehensible). There it would be helpful if one could judge a whole document as "RED" without judging every single video.	Some of the film clips were poorly edited and therefore did not match the translated text. Certain written formulations are not common in Switzerland. There are some very German formulations. The German text was taken over, there was no real translation.	The large amount of nonsense translations could lead to the fact that one does not work concentrated any more.
What were the main or most common issues with the automatic translations?			
(original in German) Es gab wenig Probleme technischer Art. Nur 1x kein Zugang zum Dokument. Ab und zu (aber selten!) eine Meldung, dass die "Resultate" nicht angenommen/gespeichert werden konnten.	(translated into English) There were few problems of a technical nature. Only 1x no access to the document. Now and then (but rarely!) a message that the "results" could not be accepted/saved.	Some of the film clips were poorly edited and therefore did not match the translated text.	The large amount of nonsense translations.

Table 10: Feedback comments from evaluators about the human evaluation setup and the Appraise platform.