Findings of the First WMT Shared Task on Sign Language Translation (WMT-SLT22)

Mathias Müller University of Zurich
Sarah Ebling University of Zurich
Eleftherios Avramidis DFKI Berlin
Alessia Battisti University of Zurich
Michèle Berger HfH Zurich
Richard Bowden University of Surrey
Annelies Braffort University of Paris-Saclay
Necati Cihan Camgöz Meta Reality Labs
Cristina España-Bonet DFKI Saarbrücken
Roman Grundkiewicz Microsoft
Zifan Jiang University of Zurich
Oscar Koller Microsoft
Amit Moryossef Bar-Ilan University
Regula Perrollaz HfH Zurich
Sabine Reinhard HfH Zurich
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 First WMT Shared Task on Sign Language Translation (WMT-SLT22)[1]. This shared task is concerned with automatic translation between signed and spoken[2] languages. The task is novel 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 featured two tracks, translating from Swiss German Sign Language (DSGS) to German and vice versa. Seven teams participated in this first 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, reproducible baseline systems and new protocols and software for human evaluation. Finally, the task also resulted in the first publicly available set of system outputs and human evaluation scores for sign language translation.

1 Introduction

This paper presents the outcome of the First WMT Shared Task on Sign Language Translation (WMT-SLT22). The focus of this shared task is automatic translation between signed and spoken languages. Recently, Yin et al. (2021) called for including signed languages in NLP research. We regard our shared task as a direct answer to this call. While WMT has a long history of shared tasks for spoken languages (Akhbardeh et al., 2021), this is the first time that signed languages are included in a WMT shared task.

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, solutions need to consider a combination of Natural Language Processing (NLP) and computer vision (CV) techniques.

In the field of sign language MT there is a general lack of suitable and freely available datasets and code. For this reason it was necessary for us to build and distribute novel training corpora and we also published reproducible baseline code. Likewise, existing protocols and toolkits for human evaluation had to be adapted to support sign languages.

In this first edition of the shared task we considered one language pair: Swiss German Sign Language (DSGS) and German. We offered two tracks: DSGS-to-German translation and German-to-DSGS translation.

Seven 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 the German-to-DSGS track.

The remainder of this paper is organized as follows:

[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 Seventh Conference on Machine Translation (WMT), pages 744 - 772

- 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 baseline 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

We consider sign language processing (SLP) a sub-area of Natural Language Processing (NLP), and automatic sign language translation (SLT), a more narrowly focused discipline within SLP.

We first 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), of SLT in particular (§2.4) and finally motivate this shared task (§2.5).

2.1 Sign languages

Sign languages (SLs) are the natural languages used in deaf communities. Contrary to the popular belief that sign language is universal, hundreds of different SLs have been documented so far. They are still scarcely described and under-resourced. For example, few reference grammars exist, lexicons only have partial coverage and existing corpora are small.

Nature of sign languages Sign languages are visuo-gestural languages. A person expresses themselves using many parts of the body (hands and arms, but also face, mouthing, gaze, shoulders, torso, etc.) while the interlocutor perceives the message through the visual channel. The linguistic system of SLs makes use of these specific linguistic cues. 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 do not have a written form or graphical system for transcription that is universally accepted (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.

A common misconception among MT researchers is that transcribed glosses are a full-flegded writing system for sign languages. In reality, glossing is a linguistic tool, useful for annotating corpora for linguistic studies (Johnston, 2010). Glosses are not a means of writing SL, and they do not adequately represent the meaning of an SL utterance. Importantly, "deaf people do not read or write glosses" in everyday life (Anonymous, 2022). Moreover, glosses mostly consist of words taken from the surrounding spoken language, which is generally only a second language to deaf signers (§2.2, societal relevance).

2.2 Relevance of sign language processing

SLP is a research area with high potential impact, as it is relevant in a societal and academic sense.

Societal relevance The overall aim of SLP is to provide language technology for sign languages, which currently are somewhat overlooked. The vast majority of NLP systems are designed for spoken languages, not for signed 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 deaf and hearing communities. 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 deaf and hearing people could benefit from such technologies in 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 can also be useful, for example to automatically

[3]We distance ourselves from the harmful 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.

subtitle signed content to make it accessible to people who do not know SLs (Bragg et al., 2019).

Academic relevance In the field of Natural Language Processing (NLP), working on SLs 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.

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 systems that transform 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 of the characteristics of sign languages (§2.1), for instance multilinearity, use of the signing space and 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 presented simultaneously, rather than sequentially. Signing makes frequent use of indexing strategies for example to identify referents introduced earlier in the discourse or timelines (Engberg-Pedersen, 1993).

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

Sign languages have an established vocabulary but are also lexically productive to allow for 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 underresourced. An analysis by Joshi et al. (2020) places all sign languages considered in this study in the category "left behind" (together with many spoken languages). Existing resources are small and also 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).

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).

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., 2020) 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 and vice versa (e.g. Camgöz et al. 2018; Yin and Read 2020), pose-to-text translation and

vice versa (e.g. Ko et al. 2019; Saunders et al. 2020a,b,c) and systems involving HamNoSys (e.g. Morrissey 2011; Walsh et al. 2022). 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 sign language corpus 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. 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.

2.5 Motivation for this shared task

Our main motivation is that sign languages are natural languages (§2.1) that are currently overlooked in NLP and SLT research (§2.3, §2.4). The shared task brings this topic to the attention of MT researchers. We decided to create a new shared task as opposed to other activities since we believe this format has a unique potential to foster progress in MT and to also make progress measurable over time.

Concrete ways in which the shared task might boost research is by creating public benchmark data, translations by many state-of-the-art systems and judgements of translation quality by humans (see also §9.4 on ways we are adding value).

3 Tracks and submission procedure

We offered two translation directions ("tracks"): translation from Swiss German Sign Language (DSGS) to German and vice versa.

Translation from DSGS to German was our primary translation direction in the sense that submitted systems were ranked on a leaderboard and we provided baseline systems. Systems translating from German to DSGS were not ranked on the leaderboard while the task was running, but we still encouraged participants to submit such systems. We were prepared to provide human evaluation for all submitted systems, regardless of the translation direction.

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 submitted translations on the OCELoT platform[5] which has a public leaderboard. We modified OCELoT slightly in order to disable automatic metrics on the leaderboard for German-to-DSGS, since currently no automatic metrics exist for SL output. Participants were allowed to make up to seven submissions, one of them the primary submission.

Main outcome Seven teams (including one from the University of Zurich 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 the second translation direction.

[5] https://ocelot-wmt22.azurewebsites.net/

		SRF		FocusNews		Total
	direction	**episodes**	**segments**	**episodes**	**segments**	**segments**
training	(both)	29	7071	197	10136	17207
development	(both)	1	287	3	133	420
test	DSGS-to-German	1	242	5	246	488
	German-to-DSGS	1	183	5	228	411

Table 1: Overview of training, development and test data. SRF and FocusNews are two different training corpora (§4.2). Segment count for the training corpora is after automatic sentence segmentation. The development data for both translation directions is identical, while the test data is different for DSGS-DE and DE-DSGS.

4 Data

For this task we provided separate training, development and test data, where the training data was available from the beginning while the development and test data were released in several stages.

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

Necessity of creating training data The data we provided are new corpora that we built and published. This was necessary because existing datasets for SL machine translation did not meet our requirements. Existing datasets either have a license that is too restrictive, are not parallel enough in the sense of being only "comparable corpora", are too small or have a very limited linguistic domain. For example, the most widely used dataset in SL machine translation research, PHOENIX (introduced in §2.4), has a size of 8k sentence pairs and contains only weather reports.

Following the long history of WMT shared tasks for spoken language machine translation (Akhbardeh et al., 2021), we opted for data that contains general news, hence a more open domain.

4.1 Licensing and attribution

Our training corpora have different licenses that are summarized here. This overview paper must be cited if the corpora are used.

FocusNews corpus This dataset can be used only for this shared task or its future iterations. Other uses of the data require express permission by the data owners. Interested parties should contact the organizers for further information.

SRF corpus This dataset can be used for non-commercial research under an Attribution-NonCommercial-ShareAlike 4.0 International license (CC BY-NC-SA 4.0)[6].

4.2 Training Data

The training data comprises two corpora, called FocusNews and SRF. The linguistic domain of both corpora is general news, and both contain parallel data between DSGS and German. The corpora are distributed through Zenodo[7].

The statistics of the two corpora are summarised in Table 2.

Training corpus 1: FocusNews [8] The FocusNews data originates from a former deaf online TV channel, FocusFive[9]. We provide the news episodes (FocusNews), as opposed to other programs. The data consists of 197 videos with associated subtitles of approximately 5 minutes each. The videos feature deaf signers of DSGS and represent the source for translation. The German subtitles were created in post-production by hearing SL interpreters.

We provide episodes within the time range of 2008 (starting with episode 43) to 2014 (up to episode 278). The videos were recorded with different framerates, either 25, 30 or 50 fps. The video resolution is 1280 x 720.

While this data set is small (by today's standards in spoken language machine translation), we emphasize the importance of using deaf signer data for shared tasks like ours. There are crucial differences between the signing of hearing interpreters and deaf signers, and interpreted signing

[6] https://creativecommons.org/licenses/by-nc-sa/4.0/

[7] https://zenodo.org/

[8] Here we describe Zenodo release version 1.3 of the corpus.

[9] https://www.youtube.com/c/focusfivetv

	FocusNews (release 1.3)	SRF (release 1.2)
Number of episodes	197	29
Time span of episodes	2008 – 2014	March 2020 – March 2021
Length of 1 episode	$\sim$ 5 minutes	$\sim$ 30 minutes
Number of signers	12	3
Signer status	deaf	hearing
Signing mode	Live signing from teleprompter (showing German text or glosses)	Live sign language interpretation
Translation source	DSGS	German
Total duration videos	19 hours	16 hours
Video resolution	1280 $\times$ 720	1280 $\times$ 720
Video framerate	25, 30 or 50	25
Number of parallel subtitles*	9943 / 10136	14265 / 7071
Number of monolingual subtitles*	(none)	883754 / 577418
Subtitle format	SRT	SRT
Sentence segmentation	automatic	manual
Subtitling mode	In post-production, after signing is already recorded	Pre-produced or live subtitles (using respeaking with ASR)

Table 2: Data statistics and characteristics of our training corpora. *= before / after automatic sentence segmentation.

may bear more resemblance to spoken language structures (Janzen, 2005).

Training corpus 2: SRF [10] The dataset contains daily national news and weather forecast episodes broadcast by the Swiss National TV (*Schweizerisches Radio und Fernsehen*, SRF)[11]. The episodes are narrated in Standard German of Switzerland (different from Standard German of Germany, and different from Swiss German dialects) and interpreted into DSGS. The interpreters are hearing individuals, some of them children of deaf adults (CODAs).

The subtitles are partly preproduced, partly created live via respeaking based on 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. This offset or "alignment shift" is visualized in Figure 1.

Manual alignment In our training corpus, the offset between the signing and the subtitles was manually corrected by deaf signers with a good command of German. The live interview and weather forecast parts of each episode were ignored, as the quality of the subtitles tends to be noticeably lower for these parts.

The parallel data comprises 29 episodes of approximately 30 minutes each with the SL videos (without audio track) and the corresponding subtitles. We selected episodes from two time spans: 13/03/2020 to 19/06/2020 and 04/01/2021 to 26/02/2021, featuring three different SL interpreters. (Three interpreters consented to having their likeness used for this shared task.) The videos have a framerate of 25 fps and a resolution of 1280 x 720.

In addition to the parallel data we provided all available German subtitles from 2014 to 2021 as monolingual data. In total, there are 1949 subtitle files with a total of 570k sentences (count after automatic segmentation).

4.3 Development data

The development data consists of segments extracted from undisclosed SRF and FocusNews episodes (see §4.2 for a general description). This data was also manually aligned and the signer is a "known" person that appeared in the training set. The framerate of development videos is 25 fps for SRF and 50 fps for FocusNews.

4.4 Test data

We distribute separate test data for our two translation directions.

[10] Here we describe Zenodo release version 1.2 of the corpus. The data provided here is an extended version of the dataset published as part of the Content4All project (EU Horizon 2020, grant agreement no. 762021).

[11] https://www.srf.ch/play/tv/sendung/tagesschau-in-gebaerdensprache?id=c40bed81-b150-0001-2b5a-1e90e100c1c0

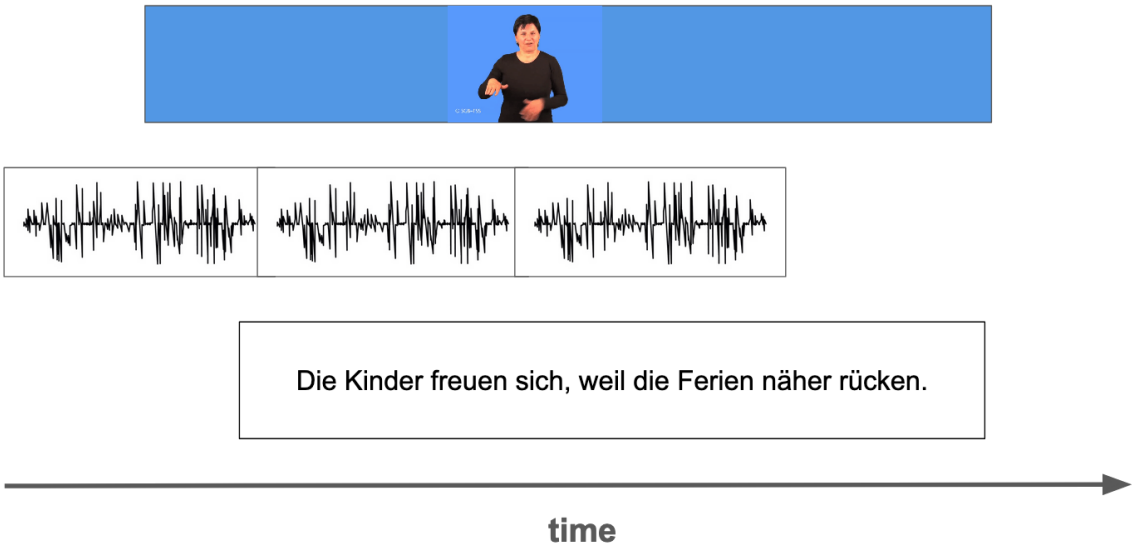

Figure 1: Illustration of alignment shift in sign language corpora. From top to bottom: a sign language video, an audio track with speech, a spoken language subtitle in German. Information in these three modalities do not start and end at the same time, adjusting their start and end times is referred to as *alignment*.

DSGS-to-German Additional, undisclosed SRF and FocusNews episodes that are manually aligned. As for the development data, the signers are "known" persons and the framerate of videos is 25 fps for SRF and 50 fps otherwise.

German-to-DSGS This subset of the test data has two distinct parts:

1. Additional, undisclosed FocusNews episodes that are manually aligned. As for the development data, the signers are "known" persons and the framerate of videos is 50 fps.

2. New translations created specifically for this shared task. The domain is identical to the training data (general news). In this case German subtitles are the source for human translation, DSGS videos are the target. The human translator is deaf (in contrast to all of the SRF data, where signers are hearing interpreters). The framerate of these videos is 50 fps and they are recorded with a green screen.

For German-to-DSGS translation we consider it important that the reference translations are created by deaf signers instead of hearing interpreters.

4.5 Automated access to training data

Our baseline system described in §6.1 automatically downloads all subsets of the data.

In addition, we added our training corpora to the Sign Language Datasets library (Moryossef and Müller, 2021b). The datasets can now be loaded automatically as a Tensorflow data set, provided that the user has previously obtained Zenodo access tokens.

5 Data preprocessing

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

5.1 Video processing

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 2 for examples.

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

Figure 2: Illustration of video preprocessing steps (cropping, instance segmentation and masking). From left to right: original frame, cropped frame, masked frame.

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).

5.2 Subtitle processing

For subtitles that are not manually aligned (all of FocusNews and monolingual SRF data), automatic sentence segmentation is used to redistribute text across subtitle segments, see Figure 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

"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)[12] and MediaPipe Holistic (Lugaresi et al., 2019)[13]. 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 3 for examples of pose estimation on our data.

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

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

OpenPose We are using the OpenPose Body135 model. 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. Values from Holistic are normalized between 0 and 1, instead of referring to actual video coordinates.

6 Baseline and submitted systems

In this section we describe all 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 six out of seven teams have given us detailed information about their submissions. The submissions are summarized in Table 4.

Overall, the participating teams have diverse academic backgrounds, most of them combine computer vision and NLP expertise. All submitted systems are sequence-to-sequence models based on Transformers (Vaswani et al., 2017). Participants chose to represent sign language data as either video frames (using a visual feature extractor on the encoder side) or pose features, with no clear majority in this regard.

Two systems, by LATTIC and MSMUNICH, are unconstrained because their visual encoder component is pretrained on WSASL (Li et al., 2020) or is an existing model taken from Varol

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.

Figure 3: Examples of the output of pose estimation systems overlaid over the original video frames. Left: OpenPose, right: MediaPipe Holistic.

	BASELINE	LATTIC	MSMUNICH	UPC	DFKI-SLT	DFKI-MLT	NJUP-MTT
Constrained	✔	-	-	✔	✔	✔	?
Multilingual	-	-	-	-	-	-	?
Document-level	-	-	-	-	-	-	?
Model ensemble	-	-	-	-	-	-	?
Pretrained components	-	✔	✔	-	-	-	?
Monolingual data	-	-	-	-	-	-	?
Synthetic data	-	-	-	-	✔	-	?
Signed language representation	OP	Video frames	Video frames	MH	MH	Video frames	?
Spoken language representation	SP	SP	other[1]	SP	other[2]	-	?
Open-source code	✔	(✔)	-	✔	✔	(✔)	?

Table 4: Overview of characteristics of submitted systems. NJUP-MTT did not disclose any information. In the code row, checkmarks are clickable links. OP=OpenPose, MH=MediaPipe Holistic, SP=Sentencepiece, (✔)=authors plan to publish the code, other[1]=text is normalized, but not segmented, other[2]=text is lowercased, but not segmented

et al. (2021). Only one team (DFKI-SLT) used synthetic parallel data and no submission used the monolingual subtitles we distributed.

Three teams have published their code, with two other teams planning to do so in the future.

6.1 Submission by UZH (baseline system)

We provided code to train baseline systems for DSGS to German in a public Github repository (Müller et al., 2022)[14]. The codebase contains scripts to preprocess data, train, translate and evaluate models and should allow to reproduce our results exactly.

The underlying sequence-to-sequence toolkit is Sockeye (Hieber et al., 2022) which is based on Pytorch (Paszke et al., 2019). We adapted Sockeye so that it supports encoding or decoding continuous vectors instead of discrete sequences of tokens. Our system is a pose-to-text translation model that reads a sequence of pose frames and converts them to the model size with a simple learned projection. The baseline does not involve pretraining or additional data and is therefore a constrained submission.

Preprocessing We used OpenPose (Cao et al., 2019) predictions (as opposed to MediaPipe Holistic or a third option). If OpenPose predicted several people in a frame, we simply chose the first one and ignored all other values. Poses are normalized by shoulder width. We convert all pose sequences to a framerate of 25 fps. On the spoken language side we do not apply any preprocessing except learning and applying a Sentencepiece segmentation model (Kudo, 2018) with a vocabulary size of 1000.

For training and translation we used one Tesla V100-32GB GPU and the training took between two and four hours.

6.2 Submission by LATTIC (Shi et al., 2022)

The system submitted by LATTIC is a Transformer-based sequence-to-sequence model which uses as input visual representations derived from an Inflated 3D ConvNet (I3D) (Carreira and Zisserman, 2017) and text as the target. The I3D models is pretrained on the WLASL[15] dataset (an isolated sign dataset). The input representation is resized video frames, the frames were resized to 224x224. For the spoken language side Sentencepiece (Kudo and Richardson, 2018) was used to generate a vocabulary of 18k tokens. The system is developed from scratch, without the use of existing MT software, and has a Transformer architecture (Vaswani et al., 2017). The I3D model is first trained on Kinetics, an action recognition dataset (Carreira and Zisserman, 2017), then it is trained for isolated sign language recognition. Before feeding input to the model, each isolated sign video is truncated, resized, randomly cropped to 224x224 and horizontally flipped with probability 0.5. Models were trained on several GPU types (A4000, A6000 and Titan RTX) and the training took roughly four hours per model.

6.3 Submission by MSMUNICH (Dey et al., 2022)

Microsoft's submission to WMT-SLT is a sequence-to-sequence Transformer model. It is based on an existing model pretrained on the

[14] https://github.com/bricksdont/sign-sockeye-baselines

[15] https://github.com/dxli94/WLASL

BSL1K dataset (Varol et al., 2021)[16]. Similar to the submission of LATTIC, this system also uses a pretrained I3D model. The system takes as input consecutive video frames and predicts over 1000 signs. For the text side, text normalisation such as lowercasing, conversion of numerals and data cleaning were applied. The authors emphasize that such careful data preprocessing and postprocessing was crucial. The underlying MT framework is Fairseq (Ott et al., 2019).

6.4 Submission by SLT-UPC (Tarrés et al., 2022)

The submission of UPC[17] is also a Transformer-based sequence-to-sequence model, based on a smaller Transformer architecture. To pretrain the model, PHOENIX (Forster et al., 2014) data was used. However, the results achieved with pretraining were no better than the primary submission (without pretraining). The authors built independent vocabularies for each training corpus. The best results were obtained by only training on the FocusNews dataset.

As a representation for the SL side, MediaPipe Holistic was used, re-extracting the features using the pose library by Moryossef and Müller (2021a). The authors interpolated the pose sequences to unify the framerate to 25fps and used data augmentation on the poses (using pose libary augmentation functions such as rotation, scaling and shear). For the text side Sentencepiece was used to generate vocabularies of 1000, 2000 and 4000. Their main submission had a vocabulary of 1000. The code is based on Fairseq and is available on GitHub[18]. To train their models, one Nvidia GeForce RTX 3090 was used and training for the main submission took roughly 3.5 hours.

6.5 Submission by DFKI-SLT (Hufe and Avramidis, 2022)

The submission of DFKI-SLT is a sequence-to-sequence model trained with JoeyNMT (Kreutzer et al., 2019), using chrF as the validation metric. The authors describe their system as having three main modules. In the first, SL images are converted into intermediate pose keypoint representations; the second module employs data augmentation (geometrical transformations) to increase sample efficiency and decrease the effect of spurious feature correlations; and the third employs a Transformer network to perform translation.

The system is trained only on FocusNews. The representation of the SL side was based on MediaPipe Holistic. The text side was only lowercased and the maximum sentence length was set to 400. The models were trained on an Nvidia RTXA6000.

6.6 Submission by DFKI-MLT (Hamidullah et al., 2022)

The main idea behind the DFKI-MLT approach is to learn feature representation and translation in a single model, and to train them together. The system architecture consists of two connected blocks: the first block, implemented using CNNs, is intended to capture visual representations and the second one, implemented with Transformers, aims to capture language. The visual component is based on a ResNet (Hara et al., 2017). In particular, the visual encoding in the submitted system consists of the original 3D ResNet10 with output conversion. The conversion creates a sequence of vectors from the single output vector to adapt to the Transformer encoder input. The visual vector is projected through a linear layer which is connected directly to the language block. The language block is a simple Transformer. The training is end-to-end, aiming to force the visual block to take into account the language representation when building the visual embedding.

6.7 Submission by NJUPT-MTT

Finally, we received submissions from the machine translation lab at Nanjing University of Posts and Telecommunications (NJUPT-MTT). 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

In our human evaluation, we followed the setting established by the recent WMT21 conference (Akhbardeh et al., 2021) and adapted it to the requirements of SLT evaluation.

[16] https://www.robots.ox.ac.uk/~vgg/research/bslattend/

[17] https://www.upc.edu/ca

[18] https://github.com/mt-upc/fairseq/tree/wmt-slt22

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), which was piloted at the IWSLT 2022 evaluation campaign (Anastasopoulos et al., 2022). 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.

Human evaluation settings We used the Appraise evaluation framework[19] (Federmann, 2018) for collecting segment-level judgements within document context. As there were submissions in the DSGS-to-German direction only (§3), 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 from an MT system. The reference translation and the official baseline were included as additional system outputs. Documents longer than ten segments were split into document snippets with ten or fewer consecutive segments. A screenshot of an example annotation in Appraise is presented in Figure 4.

We hired four evaluators who were native German speakers and trained DSGS interpreters. They did not have prior experience with evaluation of MT output. Each evaluator was assigned an identical set of annotation tasks comprising documents from the entire test set and all participating systems, including the baseline system and the reference translation. 196 segments were given to each annotator more than once to conform to Appraise's requirement of 100 segments per task and in order to measure intra-annotator agreement.

We did not include any quality control items in the annotation tasks as we had multiple independent 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.

Justification for custom guidelines We designed custom guidelines to account for different modalities (e.g. avoid confusing mentions of "text" in the instructions when the source or target is in fact a video) and to tailor them towards SL content. For example, we added *naturalness of motion* as an evaluation criterion for evaluations with SL output. Following IWSLT 2022, we also removed any mention of "grammar" to shift emphasis away from grammatical issues towards translation-breaking differences in meaning. The full instructions to evaluators in English and German are listed in Appendix A.

Data and scripts used for generating tasks and computing the final system rankings are publicly available in a Github repository.[20]

7.2 Automatic evaluation

To complement our human evaluation (which provides the main ranking) we also provide an automatic evaluation. We evaluate the submissions and the baseline system 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. We use sacreBLEU (Post, 2018) for BLEU[21] and chrF[22] and the python library for BLEURT.[23] In all cases, we estimate 95% confidence intervals via bootstrap resampling (Koehn, 2004) with 1000 samples.

8 Results

8.1 Human evaluation

Assessment scores Three out of the four evaluators completed all tasks, which gave us at least three independent judgements for each segment from the official test set. In total, for the output of eight systems, we collected 133,000 segment-level and 1,191 document-level assessment scores, which averages to 1,811.4 scores per system.

System ranking The system ranking is based on the average DA segment-level scores computed from the human assessment scores. We did not

[19] https://github.com/AppraiseDev/Appraise

[20] https://github.com/WMT-SLT/wmt-slt22

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

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

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

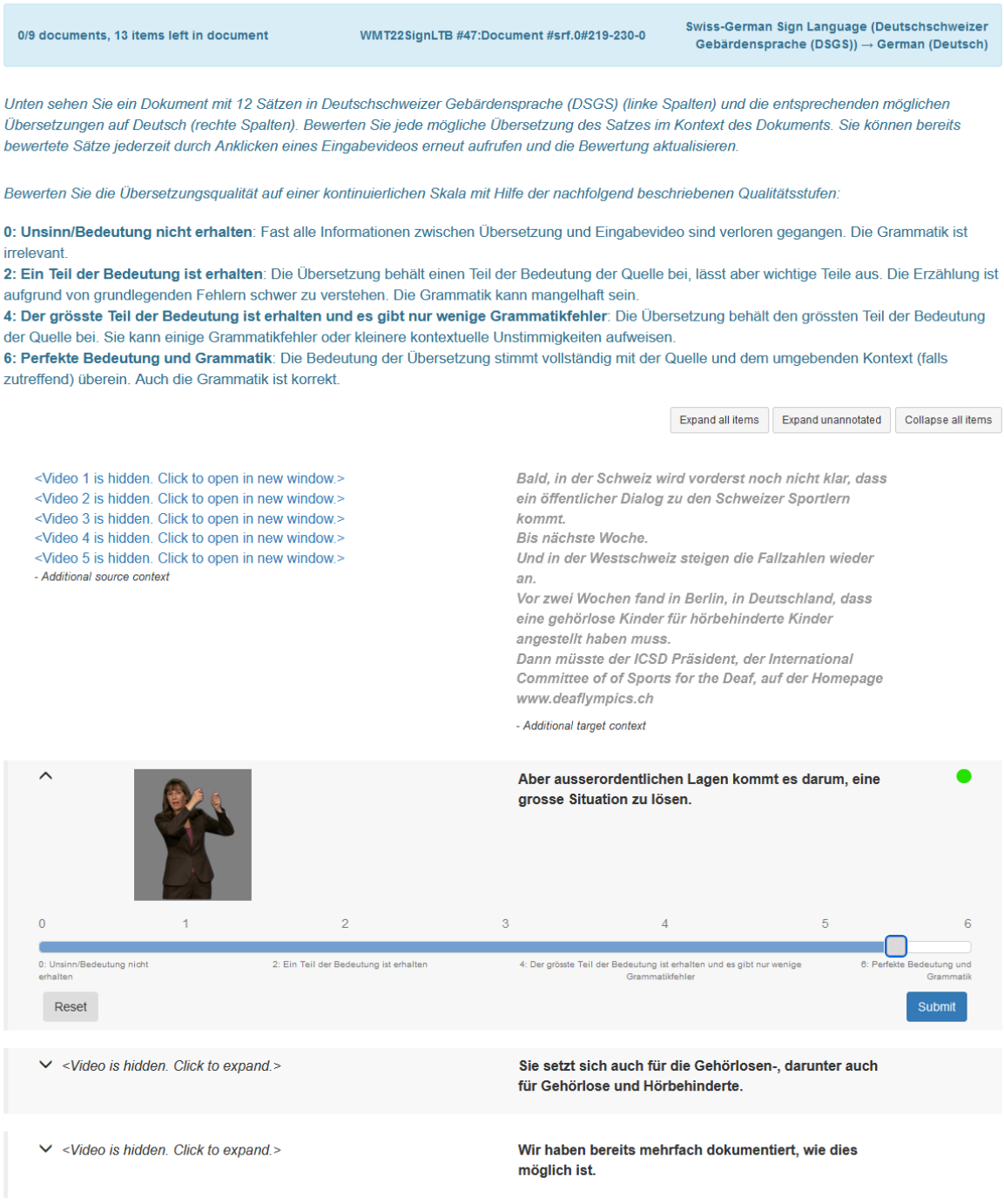

Figure 4: A screenshot of an example sign-to-text annotation task in Appraise featuring document-level source-based direct assessment (DA) with scalar quality metrics (SQM) and custom annotator guidelines in German.

make any distinction between segment-level and document-level scores, simply including the latter as additional data for computing the average scores.

The official system ranking is presented in Table 5. Systems which significantly outperform all others, according to Wilcoxon rank-sum test $p < 0.05$, are grouped into clusters, which is indicated by horizontal lines. Rank ranges giving an indication of the respective system's translation quality within a cluster are based on the same head-to-head statistical significance tests. Contrary 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.

According to our human evaluation (Table 5), MSMUNICH and LATTIC have the highest quality score among all MT systems. All other systems ended up in the same cluster with overall lower translation quality. Both winning systems are unconstrained, having been pretrained on other SL datasets, and achieve an average score of 2 in the continuous range of [0, 100], as compared to a a score of 87 for human translations and 0.52 for the baseline system. By looking at the domain-specific results, however, one can see that the performance of these two systems is around 3.5 for the FocusNews part of the test set and only 0.28-0.38 for the SRF part.

We show an additional analysis of the score distribution for each system in Appendix D.

Annotator agreement In Table 6 we are reporting intra-annotator agreement, measured with Fleiss κ (Fleiss, 1971) only as 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. One can observe that the intra-annotator agreement for raters 1 and 2 is *good* whereas for raters 3 and 4 is *very good* (Landis and Koch, 1977; Agresti, 1996).

In order to ensure the agreement between the annotators, we computed the ranks with different combinations of annotators and we did not observe changes in the ranks.

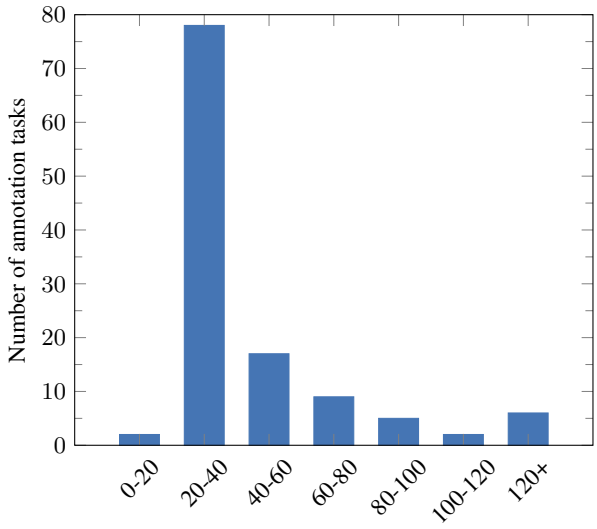

Figure 5: Number of task completion times (a task consists of 100 segments) grouped into 20-minute buckets, after removing top and bottom 5-percentiles.

Evaluation speed Three evaluators have completed the entire evaluation. A single task requiring 100 segment-level and about 12 document-level annotations took on average 45 minutes to complete, after excluding 5% of slowest and fastest task annotations. The majority of tasks were finished in between 20 and 40 minutes as shown in Figure 5.

On average, evaluators judged with a speed of 200 to 250 sentence pairs per hour. This is in line with previous evaluations for spoken language MT. We believe having such an estimate of evaluation speed is useful for future evaluations.

Feedback from evaluators After completing the evaluation two out of four 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, and that our instructions were clear. At least two people would be willing to do similar work in the future. Concerning Appraise development, at least two people experienced technical problems[24] and evaluators suggested that the user interface could be improved in some places. For instance, automatically playing videos could make evaluations more efficient.

[24]During the evaluation period there were major outages on Azure and the technical issues reported by our evaluators may be unrelated to the user interface or evaluation task.

all			SRF			FN		
Rank	Ave.	System	Rank	Ave.	System	Rank	Ave.	System
1	87.051	HUMAN	1	87.051	HUMAN	1	93.568	HUMAN
2-3	2.075	MSMUNICH	2-3	2.075	MSMUNICH	2-3	3.833	MSMUNICH
2-3	2.008	SLATTIC	2-3	2.008	SLATTIC	2-3	3.610	SLATTIC
4-5	0.520	UZH (baseline)	4-5	0.520	UZH (baseline)	4-6	1.028	UZH (baseline)
4-8	0.437	DFKI-MLT	4-8	0.437	DFKI-MLT	4-7	0.853	DFKI-MLT
5-8	0.339	DFKI-SLT	5-8	0.339	DFKI-SLT	4-7	0.671	DFKI-SLT
5-8	0.207	UPC	5-8	0.207	UPC	5-8	0.407	UPC
5-8	0.041	NJUPT-MTT	5-8	0.041	NJUPT-MTT	7-8	0.033	NJUPT-MTT

Table 5: Official results of the WMT22 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 $p < 0.05$. Gray rows indicate unconstrained systems.

annotator	κ	**items**
1	0.77±0.07	235
2	0.76±0.13	62
3	0.90±0.06	235
4	0.88±0.06	235

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

Informally, evaluators have 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. They have also pointed out that showing machine-translated target context can be confusing because for our use case quality is so low.

More detailed feedback forms submitted by evaluators are listed in Appendix C.

8.2 Automatic Evaluation

Table 7 summarises the results of the automatic evaluation. We report the scores for the full test set and also for the SRF and FocusNews subsets and boldface the primary submissions that have been evaluated manually. The low scores for all systems and metrics demonstrate the difficulty of the task. For most systems but SLATTIC with BLEU, translation quality is higher for FocusNews than for SRF. This might be an effect of the length of the source videos: SRF videos are six times longer than FocusNews, which might make the alignment with the textual part more difficult at sentence level.

The best system in the automatic evaluation depends on the evaluation metric. MSMUNICH.2 is the best system according to BLEU, SLATTIC.4 according to chrF and MSMUNICH.1 according to BLEURT. Notice that only the best system according to BLEU among these three was submitted as primary system and therefore manually evaluated. This shows that participants probably used mainly BLEU as the metric for development, except DFKI-SLT who reported that they used chrF because BLEU was always zero.

The correlation between human rankings and automatic metrics is delicate because we only have seven data points. The metric that correlates best with human scores at system level is BLEU ($r = 0.510$, $\rho = 0.571$) followed by chrF ($r = 0.508$, $\rho = 0.214$). BLEURT shows only a weak correlation with $r = 0.314$ and $\rho = 0.286$. In our scenario, translation quality is really low, and the sentences that have been properly translated are very short (e.g. *Bis nächste Woche.*). In this case, n-gram matching metrics perform better than semantic metrics.

See Appendix B for an extended discussion of the correlation between all automatic metrics (BLEU, chrF, BLEURT).

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 4 out of 100, which indicates that current automatic translations are not usable in practice, unlike spoken language MT where in specific scenarios experiments have

Submission	BLEU			chrF			BLEURT		
	all	SRF	FN	all	SRF	FN	all	SRF	FN
UZH (baseline)	0.12±0.06	0.09±0.03	0.19±0.11	5.5±0.5	5.2±0.5	5.8±0.8	0.102±0.006	0.095±0.006	0.110±0.009
DFKI-SLT	0.08±0.01	0.10±0.04	0.11±0.02	18.2±0.4	17.9±0.5	18.7±0.6	0.109±0.006	0.093±0.004	0.122±0.009
DFKI-MLT.1	0.07±0.05	0.05±0.02	0.12±0.10	6.6±0.5	6.4±0.6	6.9±0.6	0.100±0.008	0.097±0.009	0.100±0.012
DFKI-MLT.2	0.11±0.06	0.08±0.03	0.17±0.13	6.8±0.5	7.0±0.7	6.5±0.7	0.083±0.008	0.074±0.008	0.091±0.013
DFKI-MLT.3	0.08±0.04	0.06±0.02	0.13±0.10	6.5±0.5	6.8±0.8	6.2±0.7	0.075±0.009	0.067±0.009	0.081±0.014
DFKI-MLT.4	0.02±0.01	0.02±0.01	0.04±0.02	3.6±0.2	3.4±0.3	3.8±0.3	0.066±0.004	0.063±0.004	0.070±0.008
DFKI-MLT.5	0.04±0.02	0.03±0.00	0.08±0.04	5.4±0.3	5.1±0.3	5.6±0.4	0.078±0.004	0.074±0.005	0.080±0.007
MSMUNICH.1	0.44±0.21	0.34±0.18	0.63±0.35	17.1±0.5	16.3±0.7	17.8±0.9	0.166±0.013	0.147±0.012	0.179±0.022
MSMUNICH.2	0.56±0.30	0.28±0.13	0.84±0.51	17.4±0.5	17.0±0.5	17.9±0.8	0.150±0.011	0.132±0.008	0.163±0.019
NJUPT-MTT.1	0.09±0.01	0.13±0.03	0.13±0.03	14.6±0.5	14.8±0.7	14.4±0.8	0.127±0.006	0.125±0.007	0.130±0.009
NJUPT-MTT.2	0.10±0.01	0.13±0.03	0.14±0.03	14.1±0.5	14.2±0.7	14.0±0.7	0.117±0.006	0.117±0.007	0.117±0.009
SLATTIC.1	0.25±0.12	0.30±0.18	0.24±0.10	19.5±0.4	19.2±0.5	19.8±0.7	0.074±0.010	0.055±0.007	0.090±0.016
SLATTIC.2	0.20±0.14	0.32±0.23	0.10±0.02	17.9±0.5	17.4±0.7	18.5±0.8	0.092±0.012	0.080±0.010	0.098±0.017
SLATTIC.3	0.14±0.09	0.21±0.16	0.09±0.06	17.4±0.5	17.0±0.6	17.8±0.7	0.096±0.012	0.081±0.010	0.106±0.019
SLATTIC.4	0.19±0.15	0.28±0.23	0.11±0.02	19.9±0.5	19.9±0.6	19.8±0.8	0.088±0.011	0.067±0.006	0.107±0.019
SLATTIC.5	0.18±0.06	0.21±0.09	0.19±0.10	17.9±0.5	17.4±0.6	18.3±0.8	0.107±0.011	0.093±0.007	0.119±0.019
SLATTIC.6	0.07±0.03	0.15±0.07	0.04±0.01	15.0±0.4	14.8±0.5	15.0±0.6	0.103±0.010	0.094±0.006	0.110±0.017
SLT-UPC.1	0.34±0.22	0.29±0.14	0.43±0.33	15.6±0.6	15.4±0.8	15.8±0.8	0.131±0.005	0.126±0.006	0.136±0.008
SLT-UPC.2	0.35±0.21	0.29±0.14	0.43±0.30	16.2±0.6	15.4±0.8	17.0±0.9	0.136±0.004	0.126±0.006	0.145±0.007
SLT-UPC.3	0.41±0.33	0.24±0.10	0.54±0.47	15.5±0.6	15.1±0.8	16.0±0.9	0.144±0.006	0.131±0.006	0.157±0.010
SLT-UPC.4	0.28±0.09	0.26±0.11	0.37±0.16	12.2±0.4	12.3±0.6	12.1±0.6	0.113±0.004	0.122±0.006	0.103±0.006
SLT-UPC.5	0.24±0.10	0.32±0.14	0.25±0.12	12.0±0.4	12.1±0.6	11.9±0.5	0.102±0.004	0.110±0.006	0.094±0.006
SLT-UPC.6	0.28±0.09	0.26±0.11	0.37±0.16	12.2±0.4	12.3±0.6	12.1±0.6	0.113±0.004	0.122±0.006	0.103±0.006
SLT-UPC.7	0.50±0.26	0.37±0.13	0.61±0.38	12.3±0.5	11.9±0.7	12.7±0.8	0.111±0.006	0.110±0.007	0.111±0.011

Table 7: Automatic evaluation of all the submission for the full WMT-SLT test set (all), the SRF subset and the FocusNews (FN) subset. Mean and 95% confidence intervals obtained via bootstrap resampling are shown. Primary submissions manually evaluated are boldfaced. Note that the official ranking is given by the human evaluation (Table 5).

shown systems to be on par with human translation (Hassan et al., 2018; Popel et al., 2020). In the following paragraphs we discuss potential reasons for this outcome.

Size of training data The corpora we have built for this shared task (§4) are superior to existing datasets (in terms of size, license, linguistic domain and alignment quality), but are still small. Taken together our corpora contain 20k parallel sentence pairs only, and 600k monolingual German sentences. This limits the optimal translation quality that could in theory be obtained in a constrained setup. This is corroborated by the fact that the two unconstrained systems have won the shared task (§8).

Building larger parallel SL corpora in itself is challenging. Even though recently steps were taken to collect larger amounts of data (e.g. in the projects EASIER and SignON), such resources are not immediately useful because basic linguistic tools used to prepare parallel corpora are not available (§2.3, lack of basic linguistic tools). For spoken language NLP, such tools are commonplace, work well and are used to automatically compile large corpora. For example, Bitextor[25], a tool developed in the Paracrawl project (Bañón et al., 2020), relies on the automatic alignment tool BleuAlign (Sennrich and Volk, 2011).

Modality gap But even if much more training data was available, it is likely that current MT methods are not adapted well enough to SL data. NLP methods in general are tailored towards text and may perform worse or not be applicable at all to other modalities. For example, there are currently no efficient tools for automatic SL segmentation (Yin et al., 2021), while for text-based MT, subword segmentation (Sennrich et al., 2016; Kudo, 2018) has become a staple in research.

While all systems submitted this year are signed-to-spoken systems, the modality gap is more apparent for automatic spoken-to-sign translation because generating continuous outputs requires more fundamental changes to existing MT toolkits (as opposed to the changes necessary for continuous inputs).

[25] https://github.com/bitextor/bitextor

The proclivity of existing MT research for text data is confirmed by the number of recent works that chose to represent SL content as (textual) gloss sequences, despite the fact that glosses are not an adequate representation of meaning (Anonymous, 2022).

9.2 Reliability of evaluation procedure

Our evaluation is reliable since we conduct a human evaluation (compared to other shared tasks which produce official rankings based on automatic metrics). But even compared to shared tasks that do offer human evaluation (such as the General task this year), we believe that our evaluation is strong, since we have three to four (at least three) independent judgements for each system output across the entire test set.

9.3 Limitations of shared task setup

We note several limitations of the specific experimental setup in this year's shared task.

Generalization As explained in §4 all signers that appear in the development and test sets are known, in the sense of also being present in the training data. It is therefore important to emphasize that our shared task evaluates the performance of systems on familiar signers, and does not test generalization to unseen individuals.

Recording conditions Since our training data is derived from news broadcasts, the recording conditions and video quality are favourable. For example, the signer is always recorded against a monochrome and static background. The recording angle is very consistent, as cameras are mounted on a fixed rig. Signers always directly face the camera. The recording conditions therefore resemble laboratory conditions.

This means that our shared task does not evaluate "signing in the wild" (examples: mobile recordings of varying quality, varying angles, moving background including other people) and it is likely that the outcome would be different in that case.

Interpretation vs. translation Some of our training material is interpreted live (§4). Interpretation has constraints that are very different from offline translation, most notably, interpreters are under severe time pressure. This has consequences for the resulting signed material, which may sometimes omit phrases to keep up with the narrative, or interpreters would sign an utterance differently if they could give it a second thought.

A general property of SL interpretation (and hearing signers in general, as opposed to deaf signers) is that its linguistic structure tends to follow the structure imposed by the spoken language being translated (Janzen, 2005). This means that systems trained on such material may resemble hearing interpreters more than deaf translators.

9.4 Value created by this shared task

This shared task provides new insights and resources that previously did not exist for SLT, and that are valuable for the community.

We provided new training corpora and an official development and test set. We open-sourced a baseline system and code that is fully reproducible. We design protocols for human evaluation and adapt existing evaluation software accordingly. Lastly, the shared task resulted in the first openly available set of human judgements of automatic SL translations. Future work could use these scores for metric development, for instance.

10 Conclusion and future directions

In this paper we present the first WMT Shared Task on Sign Language Translation (WMT-SLT22). 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 (§2).

Seven teams participated in this first edition of the shared task. Overall, we observed low system performance with an average human evaluation score of about 4 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 Future iterations of the shared task could introduce more language pairs and larger training data. Since this year all submissions are signed-to-spoken systems, the shared task could also focus more on sign language generation going forward.

Furthermore, we will consider introducing additional MT-related tasks such as a sign language version of the metrics task. This perhaps requires a better distribution of human evaluation scores,

as our current set of scores very much focuses on both ends of the score spectrum (we do not have many mid-range scores).

Finally, future human evaluation experiments for spoken-to-signed translation could be run differently than explained in this paper. Namely, for campaigns where a sign language is the target language the evaluation could be reference-based instead of source-based. The advantage of this change would be that deaf evaluators can perform this evaluation, instead of hearing interpreters for whom in this case the target language is not their first language.

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 individuals and should avoid claiming that their approach "solves" any particular problem. Ideally, research of this nature should include deaf people, not only at evaluation time, but in the entire development cycle.

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) video features outperform pose features.

Acknowledgments

This shared task was funded by the European Association for Machine Translation (EAMT) and by Microsoft AI for Accessibility. We are grateful for their support which enabled us to provide test data, human evaluation and interpretation in International Sign during the WMT conference.

The organizing committee further acknowledge 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 through the project SocialWear (01IW20002).

Thanks to Stanko Pavlica from FocusFilm and Robin Ribback from SWISSTXT for their help with data and licensing. Thanks to Tom Kocmi and Christian Federmann from Microsoft Munich for their help with the submission and evaluation platform. Thanks to Yvette Graham, Nitika Mathur and Tom Kocmi for providing expertise on statistical analysis.

Finally we would like to extend heartfelt thanks to the DSGS interpreters who performed our human evaluation: Michèle Berger, Sarah Caminada, 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.

Antonios Anastasopoulos, Loïc Barrault, Luisa Bentivogli, Marcely Zanon Boito, Ondřej Bojar, Roldano Cattoni, Anna Currey, Georgiana Dinu, Kevin Duh, Maha Elbayad, Clara Emmanuel, Yannick Estève, Marcello Federico, Christian Federmann, Souhir Gahbiche, Hongyu Gong, Roman Grundkiewicz, Barry Haddow, Benjamin Hsu, Dávid Javorský, Věra Kloudová, Surafel Lakew, Xutai Ma, Prashant Mathur, Paul McNamee, Kenton Murray, Maria Nădejde, Satoshi Nakamura, Matteo Negri, Jan Niehues, Xing Niu, John Ortega, Juan Pino, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Marco

Turchi, Yogesh Virkar, Alexander Waibel, Changhan Wang, and Shinji Watanabe. 2022. Findings of the IWSLT 2022 Evaluation Campaign. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 98–157, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Anonymous. 2022. Considerations for meaningful sign language machine translation based on glosses. Anonymous preprint under review.

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.

Claudia Bianchini and Fabrizio Borgia. 2012. Writing sign languages: analysis of the evolution of the signwriting 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, Michèle Gouiffès, and Annelies Braffort. 2020. 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*.

João Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733.

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.

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.

Subhadeep Dey, Abhilash Pal, Cyrine Chaabani, and Oscar Koller. 2022. Clean Text and Full-Body Transformer: Microsoft's Submission to the WMT22 Shared Task on Sign Language Translation. In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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).

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.

Yasser Hamidullah, Josef van Genabith, and Cristina España-Bonet. 2022. DFKI-MLT at WMT-SLT22: Spatio-temporal Sign Language Representation and Translation. In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi, United Arab Emirates. 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.

Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. 2017. Learning spatio-temporal features with 3d residual networks for action recognition. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 3154–3160.

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*.

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

Lorenz Hufe and Eleftherios Avramidis. 2022. Experimental Machine Translation of the Swiss German Sign Language via 3D augmentation of body keypoints. In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Terry Janzen. 2005. *Topics in signed language interpreting: Theory and practice*, volume 63. John Benjamins Publishing.

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.

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, 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.

Julia Kreutzer, Jasmijn Bastings, and Stefan Riezler. 2019. Joey NMT: A Minimalist NMT Toolkit for Novices. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 109–114, Hong Kong, China. Association for Computational Linguistics.

Taku Kudo. 2018. Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

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.

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.

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. 2021a. pose-format: Library for viewing, augmenting, and handling .pose files. https://github.com/AmitMY/pose-format.

Amit Moryossef and Mathias Müller. 2021b. Sign Language Datasets. https://github.com/sign-language-processing/datasets.

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, Annette Rios, and Amit Moryossef. 2022. Sockeye baseline models for sign language translation. https://github.com/bricksdont/sign-sockeye-baselines.

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.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

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.

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.

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, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich and Martin Volk. 2011. Iterative, MT-based Sentence Alignment of Parallel Texts. In *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011)*, pages 175–182, Riga, Latvia. Northern European Association for Language Technology (NEALT).

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. TTIC's WMT-SLT 22 Sign Language Translation System. In *Proceedings of the Seventh Conference on Machine Translation*, 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.

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

Laia Tarrés, Gerard I. Gállego, Xavier Giró i Nieto, and Jordi Torres. 2022. Tackling Low-Resource Sign Language Translation: UPC at WMT-SLT 22. In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

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.

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.

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.

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 Appraise instructions to human evaluators

A.1 Sign-to-text direction

A.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.

A.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.

A.2 Text-to-sign direction

A.2.1 English

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

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

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

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. Naturalness of motion is irrelevant.
- 2: Some Meaning Preserved: The translation preserves some of the meaning of the source but misses significant parts. The narrative is hard to follow due to fundamental errors. Naturalness of motion may be poor.

- 4: Most Meaning Preserved and Acceptable Natural Motion: The translation retains most of the meaning of the source. It may have some minor mistakes or contextual inconsistencies. Motion may appear unnatural.
- 6: Perfect Meaning and Naturalness: The meaning of the translation is completely consistent with the source. Motion is natural.

A.2.2 German

Unten sehen Sie ein Dokument mit 10 Sätzen auf Deutsch (linke Spalten) und die entsprechenden möglichen Übersetzungen in Deutschschweizer Gebärdensprache (DSGS) (rechte Spalten). Bewerten Sie jede mögliche Übersetzung des Satzes im Kontext des Dokuments. Sie können bereits bewertete Sätze jederzeit durch Anklicken eines Quelltextes 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 Ausgangstext sind verloren gegangen. Es ist irrelevant, ob die Bewegungen natürlich sind.
- 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. Bewegungen können mangelhaft sein.
- 4: Der grösste Teil der Bedeutung ist erhalten und die Bewegungen sind akzeptabel: Die Übersetzung behält den grössten Teil der Bedeutung der Quelle bei. Sie kann kleine Fehler oder kleinere kontextuelle Unstimmigkeiten aufweisen. Bewegungen sehen teilweise nicht natürlich aus.
- 6: Perfekte Bedeutung und Natürlichkeit: Die Bedeutung der Übersetzung stimmt vollständig mit der Quelle und dem umgebenden Kontext (falls zutreffend) überein. Bewegungen wirken natürlich.

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 Ausgangstext sind verloren gegangen. Es ist irrelevant, ob die Bewegungen natürlich sind.
- 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. Bewegungen können mangelhaft sein.
- 4: Der grösste Teil der Bedeutung ist erhalten und die Bewegungen sind akzeptabel: Die Übersetzung behält den grössten Teil der Bedeutung der Quelle bei. Sie kann kleine Fehler oder kleinere kontextuelle Unstimmigkeiten aufweisen. Bewegungen sehen teilweise nicht natürlich aus.
- 6: Perfekte Bedeutung und Natürlichkeit: Die Bedeutung der Übersetzung stimmt vollständig mit der Quelle und dem umgebenden Kontext (falls zutreffend) überein. Bewegungen wirken natürlich.

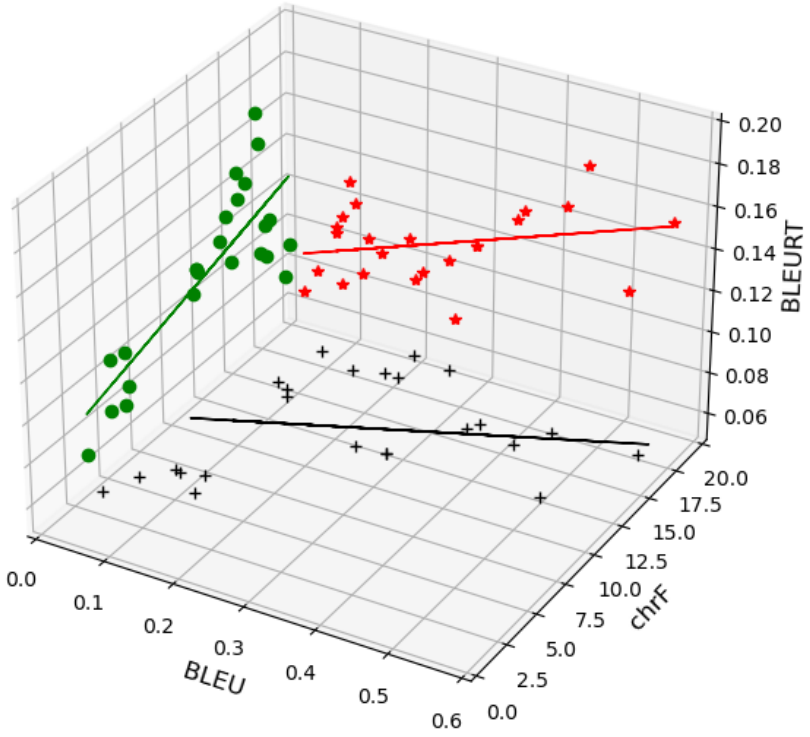

Figure 6: Correlation between the metrics used in the automatic evaluation. Automatic evaluation scores projected into the 2D spaces for BLEU–chrF (black crosses, $r = 0.447$), BLEU–BLEURT (red stars, $r = 0.703$) and chrF–BLEURT (green dots, $r = 0.443$).

B Correlation between automatic metrics

Metrics do not correlate well with each other, especially if chrF is compared to a second metric. Figure 6 plots the projection for the scores on the full test set by metric pair for the 23 submissions and the baseline. The Pearson correlation shows that metrics are far from a linear relation: BLEU–chrF has $r = 0.447$, BLEU–BLEURT $r = 0.703$ and chrF–BLEURT $r = 0.443$. Spearman correlation, accounting only for monotonicity, is lower in the three cases specially for chrF–BLEURT ($\rho = 0.259$), with $\rho = 0.421$ for BLEU–chrF and $\rho = 0.633$ for BLEU–BLEURT.

C Feedback from evaluators

Table 8 lists detailed by evaluators regarding the human evaluation procedure and the Appraise system. Two out of four evaluators submitted a response.

	Answer 1	Answer 2
What is your experience in assessing machine translation outputs?		
	None: this was my first time	Low: I have done it once 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	Strongly agree
Instructions were clear	Agree	Strongly agree
Quality levels 0-6 were helpful to me	Agree	Strongly Agree
Source videos/texts were understandable	Neutral	Strongly Agree
There was too much repetitiveness	Disagree	Agree
Documents were too long	Disagree	Strongly Disagree
Segments were too short	Neutral	Disagree
In some cases, the context was insufficient	Strongly Agree	Disagree
I experienced technical issues	Agree	Agree
I would be willing to do similar work in future	Strongly agree	Agree
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?		
	it would be very helpful, if the video started automatically when moving to the next segment. (some did, but many more did not) It would save a click. Also, the submit button could be on the left side under the 0 score (at the moment, as most translation are not yet good quality)	-
What were the main or most common issues with the automatic translations?		
	This question is not clear to me. You mean on a technical level or something else? meaning was garbage, some did not know the German Umlaute äüö	-
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	2	2
stressful/relaxed	2	-2
laborious/effortless	1	2
slow/fast	0	0
inefficient/efficient	2	2
boring/exciting	-2	3
complicated/simple	2	2
annoying/enjoyable	0	1
limiting/creative	-2	0
impractical/practical	2	0

Table 8: Feedback from evaluators about the human evaluation setup and the Appraise platform.

D Human evaluation score distribution

To complement our analysis we show the distribution of scores for each system in Figure 7. The set of scores (excluding zero scores, which are not shown in the figure) resembles a bimodal distribution, with most of the scores residing at both ends of the spectrum. MSMUNICH is the system with the most scores in the highest-quality bucket.

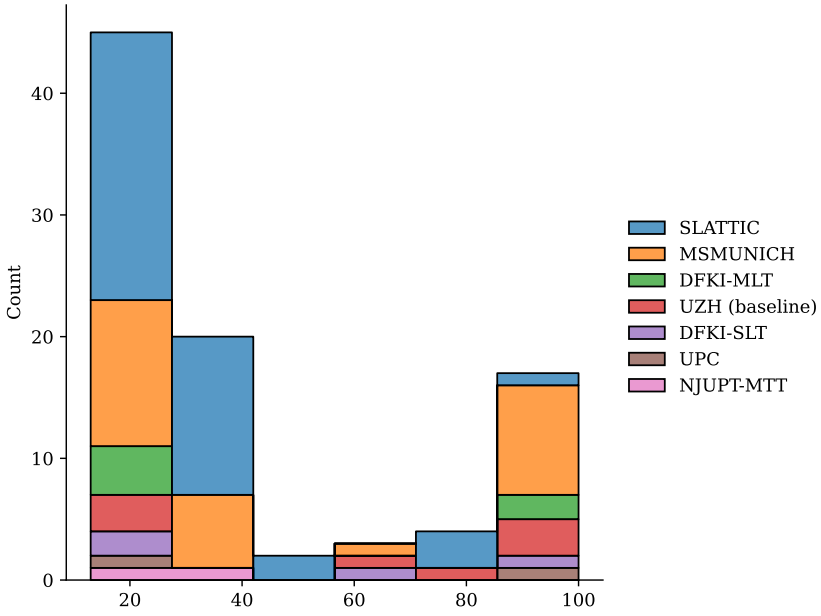

Figure 7: Distribution of human evaluation scores for all submitted systems discretized in seven bins, excluding scores of bin 0 (lowest quality).