FINDINGS OF THE IWSLT 2022 EVALUATION CAMPAIGN

Antonios Anastasopoulos George Mason U.
Loïc Barrault Le Mans University
Luisa Bentivogli FBK
Marcely Zanon Boito U. Avignon
Ondřej Bojar Charles U.
Roldano Cattoni FBK
Anna Currey AWS
Georgiana Dinu AWS
Kevin Duh JHU
Maha Elbayad Meta
Clara Emmanuel Apple
Yannick Estève Avignon University
Marcello Federico AWS
Christian Federmann Microsoft
Souhir Gahbiche Airbus
Hongyu Gong Meta
Roman Grundkiewicz Microsoft
Barry Haddow U. of Edinburgh
Benjamin Hsu AWS
Dávid Javorský Charles U.
Věra Kloudová Charles U.
Surafel M. Lakew AWS
Xutai Ma JHU/Meta
Prashant Mathur AWS
Paul McNamee JHU
Kenton Murray JHU
Maria Nădejde AWS
Satoshi Nakamura NAIST
Matteo Negri FBK
Jan Niehues KIT
Xing Niu AWS
John Ortega Le Mans University
Juan Pino Meta
Elizabeth Salesky JHU
Jiatong Shi CMU
Matthias Sperber Apple
Sebastian Stüker Zoom
Katsuhito Sudoh NAIST
Marco Turchi FBK
Yogesh Virkar AWS
Alex Waibel CMU/KIT
Changhan Wang Meta
Shinji Watanabe CMU

Abstract

The evaluation campaign of the 19th International Conference on Spoken Language Translation featured eight shared tasks: (i) Simultaneous speech translation, (ii) Offline speech translation, (iii) Speech to speech translation, (iv) Low-resource speech translation, (v) Multilingual speech translation, (vi) Dialect speech translation, (vii) Formality control for speech translation, (viii) Isometric speech translation. A total of 27 teams participated in at least one of the shared tasks. This paper details, for each shared task, the purpose of the task, the data that were released, the evaluation metrics that were applied, the submissions that were received and the results that were achieved.

1 Introduction

The International Conference on Spoken Language Translation (IWSLT) is the premier annual scientific conference for all aspects of spoken language translation. IWSLT is organized by the Special Interest Group on Spoken Language Translation, which is supported by ACL, ISCA and ELRA. Like in all previous editions (Akiba et al., 2004; Eck and Hori, 2005; Paul, 2006; Fordyce, 2007; Paul, 2008, 2009; Paul et al., 2010; Federico et al., 2011, 2012; Cettolo et al., 2013, 2014, 2015, 2016, 2017; Niehues et al., 2018, 2019; Ansari et al., 2020; Anastasopoulos et al., 2021), this year's conference was preceded by an evaluation campaign featuring shared tasks addressing scientific challenges in spoken language translation.

This paper reports on the 2022 IWSLT Evaluation Campaign, which offered eight shared tasks:

- **Simultaneous speech translation**, addressing low latency speech translation either streamed by a speech recognition (ASR) system or directly from the audio source. The translation directions for both conditions are: English to German, English to Japanese, and English to Mandarin Chinese.

- **Offline speech translation**, proposing speech

Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022), pages 98 - 158

Team	Organization
AISP-SJTU	Aispeech and Shanghai Jiao Tong University, China (Zhu et al., 2022)
ALEXA AI	Amazon Alexa AI, USA (Shanbhogue et al., 2022)
ALEXA AI	Amazon Alexa AI, USA (Zhang et al., 2022a)
APPTEK	AppTek, Germany (Wilken and Matusov, 2022)
CMU	Carnegie Mellon University, USA (Yan et al., 2022)
CUNI-KIT	Charles University, Czech Republic, and KIT, Germany (Polák et al., 2022)
FBK	Fondazione Bruno Kessler, Italy (Gaido et al., 2022)
GMU	George Mason University, USA
HW-TSC	Huawei Translation Services Center, China (Li et al.; Wang et al.; Guo et al.; Li et al.)
JHU	Johns Hopkins University, USA (Yang et al., 2022)
KIT	Karlsruhe Institute of Technology, Germany (Pham et al., 2022; Polák et al., 2022)
MLLP-VRAIN	Universitat Politècnica de València, Spain (Iranzo-Sánchez et al., 2022)
NA	Neural.AI, China
NAIST	Nara Institute of Science and Technology, Japan (Fukuda et al., 2022)
NIUTRANS	NiuTrans, China (Zhang et al., 2022c)
NUV	Navrachana University, India (Bhatnagar et al., 2022)
NEMO	NVIDIA NeMo, USA(Hrinchuk et al., 2022)
ON-TRAC	ON-TRAC Consortium, France (Boito et al., 2022b)
UOS	University of Sheffield, UK (Vincent et al., 2022)
TALTECH	Tallinn University of Technology, Estonia
UMD	University of Maryland, USA (Rippeth et al., 2022)
UPC	Universitat Politècnica de Catalunya, Spain (Tsiamas et al., 2022a)
USTC-NELSLIP	University of Science and Technology of China (Zhang et al., 2022b)
XIAOMI	Xiaomi AI Lab, China (Guo et al., 2022a)
YI	Yi, China (Zhang and Ao, 2022)

Table 1: List of Participants

translation of talks from English to German, English to Japanese, and English to Mandarin Chinese, using either cascade architectures or end-to-end models able to directly translate source speech into target text;

- **Speech to speech translation**, investigating for the first time automatic translation of human speech in English into synthetic speech in German, either with cascaded or direct neural models.

- **Low-resource speech translation**, focusing on resource-scarce settings for translating input speech in Tamasheq into French text, and input speech in Tunisian Arabic into English text.

- **Multilingual speech translation**, analyzing the performance of multi-lingual versus bilingual translation models for the Offline speech translation tasks (discussed in the Offline task section);

- **Dialect speech translation**, addressing speech translation from Tunisian into English under three training data conditions: (i) only with limited dialect-specific training data (provided by the organizers); (ii) with also larger amount of related-language data (Modern Standard Arabic); (iii) with any kind of publicly available data.

- **Formality control for SLT**, addressing the formality level (formal vs. informal) in spoken language translation from English into German, Spanish, Hindi, Japanese, Italian and Russian. The task focuses in particular on zero-shot learning in multilingual models, given that for the last two directions no formality-annotated training data is provided.

- **Isometric SLT**, addressing the generation of translations similar in length to the source, from English into French, German and Spanish.

The shared tasks attracted 27 participants (see Table 1) from both academic and industrial organizations. The following sections report on each shared task in detail, in particular: the goal and automatic metrics adopted for the task, the data used for training and testing data, the received submissions and the summary of results. Detailed results for some of the shared tasks are reported in a corresponding appendix.

2 Simultaneous Speech Translation

Simultaneous translation is the task of generating translations incrementally given partial text or speech input only. Such capability enables multilingual live communication and access to multilingual multimedia content in real time. The goal of this challenge, organized for the third consecutive year, is to examine systems that translate text or audio in a source language into text in a target language from the perspective of both translation quality and latency.

2.1 Challenge

Participants were given two parallel tracks to enter and encouraged to enter all tracks:

- text-to-text: translating the output of a streaming ASR system in real time from English to German, English to Japanese, and English to Mandarin Chinese.
- speech-to-text: translating speech into text in real time from English to German, English to Japanese, and English to Mandarin Chinese.

For the speech-to-text track, participants were encouraged to submit systems either based on cascaded or end-to-end approaches. Participants were required to upload their system as a Docker image so that it could be evaluated by the organizers in a controlled environment. We also provided example implementations and baseline systems for English-German speech-to-text translation, English-Japanese speech-to-text translation and English-Japanese text-to-text translation.

2.2 Data and Metrics

The training and development data conditions were identical as in the Offline Speech Translation track. More details are available in §3.2.

Systems were evaluated with respect to quality and latency. Quality was evaluated with the standard BLEU metric (Papineni et al., 2002) and, as a first trial this year, also manually. Latency was evaluated with metrics developed for simultaneous machine translation, including average proportion (AP), average lagging (AL) and differentiable average lagging (DAL, Cherry and Foster 2019), and later extended to the task of simultaneous speech translation (Ma et al., 2020b).

The evaluation was run with the SIMULEVAL toolkit (Ma et al., 2020a). For the latency measurement of all systems, we contrasted computation-aware and non computation-aware latency metrics. Computation-aware latency was also computed for text-to-text systems by taking into account the timestamps obtained from the ASR transcript generated by a streaming ASR model. The latency was calculated at the word level for English-German systems and at the character level for English-Japanese and English-Mandarin systems. BLEU was computed via sacrebleu (Post, 2018) (as integrated into SIMULEVAL) with default options for English-German, with the ”zh” option for English-Mandarin and with the MeCab tokenizer for English-Japanese.

The systems were ranked by the translation quality (measured by BLEU) in different latency regimes, low, medium and high. Each regime was determined by a maximum latency threshold measured by AL on the Must-C tst-COMMON set. The thresholds were set to 1000, 2000 and 4000 for English-German, 2500, 4000 and 5000 for English-Japanese and 2000, 3000 and 4000 for English-Japanese, and were calibrated by the baseline system. Participants were asked to submit at least one system per latency regime and were encouraged to submit multiple systems for each regime in order to provide more data points for latency-quality trade-off analyses. The organizers confirmed the latency regime by rerunning the systems on the tst-COMMON set.

The systems were run on the test set segmented in three ways: the first segmentation, called gold, leverages the transcript to force align and segment the audio; the second and third segmentations, called Segmentation 1 and Segmentation 2, use a voice activity detection tool to segment the input audio without relying on the transcript.

2.3 Novelties for the Third Edition

Text-to-text track moving closer to the speech-to-text track This year, we used the output of a streaming ASR system as input instead of the

gold transcript. As a result, both text-to-text and speech-to-text systems can be ranked together for a given language pair.

Language pairs We added Mandarin Chinese as a target language, resulting in three pairs: English-German, English-Japanese and English-Mandarin.

Human Evaluation and Human Interpretation Benchmark We added an experimental manual evaluation for the English-to-German speech-to-text track as well as a human interpretation benchmark (Section 2.6.1). Independently, English-to-Japanese speech-to-text track outputs were also manually scored, using the MQM setup, see Section 2.6.2.

Segmentation We reverted to the setting of the first edition where we only used segmented input in order to reduce the number of conditions and also because we noticed that existing latency metrics were not well adapted to long unsegmented input. However, recent improvements to the latency metrics (Iranzo-Sánchez et al., 2021) could allow to work with unsegmented input in the future.

2.4 Submissions

The simultaneous task received submissions from 7 teams, the highest number to date. 5 teams entered the English-German speech-to-text track, 3 teams entered the English-Mandarin speech-to-text track and 3 teams entered the English-Japanese speech-to-text track. For text-to-text, there were 3 teams for English-Mandarin, 1 team for English-German and 1 team for English-Japanese. Given that the majority of submissions were on the speech-to-text track, we are considering consolidating the task into speech-to-text only in future editions.

XIAOMI (Guo et al., 2022a) entered the text-to-text track for English-Mandarin. Their model is transformer-based and leverages R-Drop and a deep architecture. Data augmentation methods include tagged backtranslation, knowledge distillation and iterative backtranslation. Simultaneous models use the multi-path wait-k algorithm. Finally, two error correction models are introduced in order to make the systems more robust to ASR errors.

MLLP-VRAIN (Iranzo-Sánchez et al., 2022) entered the speech-to-text track for English-German. They adopt a cascaded approach, with a chunking-based DNN-HMM ASR model, followed by a multi-path wait-k transformer-based MT model. Speculative beam search is employed at inference time.

HW-TSC (Wang et al., 2022) entered all tracks, i.e. speech-to-text and text-to-text for English-German, English-Japanese and English-Mandarin. Moreover, the authors contrasted cascaded and end-to-end methods for the speech-to-text track.

CUNI-KIT (Polák et al., 2022) entered the speech-to-text track for English-German, English-Japanese and English-Mandarin. They propose a method for converting an offline model to a simultaneous model without adding modifications to the original model. The offline model is an end-to-end multilingual speech-to-text model that leverages a pretrained wav2vec 2.0 encoder and a pretrained mBART decoder. The input is broken down into chunks and decoding is run for each new chunk. Once a stable hypothesis is identified, that hypothesis is displayed. Various stable hypothesis detection methods are investigated.

AISP-SJTU (Zhu et al., 2022) entered the speech-to-text and text-to-text tracks for English-Mandarin. Their model is based on an ASR + MT cascade. They propose dynamic-CAAT, an improvement over CAAT (Liu et al., 2021) that uses multiple right context window sizes during training. The proposed method is compared to wait-k and multi-path wait-k. Data augmentation methods include knowledge distillation, tagged backtranslation and marking data with lowercased and non punctuated input with a special token.

FBK (Gaido et al., 2022) entered the speech-to-text track for English-German with an end-to-end model. The authors' main goal is to reduce computation requirements in order to democratize the task to more academic participants. First, they show how to avoid ASR encoder pretraining by using a conformer architecture and a CTC loss on top of an intermediate layer in the encoder. In addition, they use the same model for the offline task as for the simultaneous task. The auxiliary CTC loss is used to predict word boundaries and informs a wait-k policy. The latency is also controlled by the speech segment size. Finally, two data filtering methods based on negative log likelihood of an initial model and length ratio are investigated in order to make training more efficient.

NAIST (Fukuda et al., 2022) entered the speech-to-text track for English-German and English-Japanese. The proposed model applies decoding each time a new input speech segment is detected and to constrain the decoder on previously output predictions. An offline model is trained first and then finetuned on prefix pairs. The prefix pairs are extracted by translating prefixes and checking that the generated target is a prefix of the translation of the entire input. Prefixes with length imbalance are filtered out. An input segment boundary predictor is trained as a classifier by considering all prefixes and giving a positive labels to those prefixes that were extracted previously.

2.5 Results

Results are summarized in Figure 1, Figure 2 and Figure 3. We also present the text-to-text results on English-Mandarin[1] in Figure 4. More details are available in the appendix. The results include both text-to-text systems and speech-to-text systems. When participants submitted both a text-to-text system and a speech-to-text system, we retain the best system. The only participant with only a text-to-text system is XIAOMI and we can see that the system is at a disadvantage due to the noise introduced by the provided streaming ASR model. The ranking are consistent across the medium and high latency regime. However, for the low latency regime, we note a degradation from the FBK system and we observe that the NAIST system is robust to lower latency.

2.6 Human Evaluation

We conducted a human evaluation for English-to-German and English-to-Japanese independently.

2.6.1 English-to-German

For English-to-German, the human evaluation was inspired by Javorský et al. (2022). This evaluation examined (1) the best system from each latency regime selected by BLEU score, and (2) transcription of human interpretation by a professional English-German interpreter (German native speaker, certified German conference interpreter and sworn translator and interpreter) in February 2022. The interpreting was carried out remotely and transcribed by students of German for Intercultural Communication at the Institute of Translation Studies, Charles University, Faculty of Arts.[2]

The English-to-German task used two parts of the test set: (1) the Common part is used as the blind test set in the automatic evaluation and also in the Offline speech translation task, and (2) the Non-Native part comes from IWSLT 2019 Non-Native Translation Task.

Details of the human evaluation are provided in Section A.1.1 of the Appendix and results are shown in Table 18. BLEU scores correlate very well with the human judgements for both parts of the test set, as can be seen in Figure 5.

The Common part of the test set is kept confidential for future use. For the Non-Native part, we release system outputs as well as manual judgements on the corresponding IWSLT page.[3]

2.6.2 English-to-Japanese

For English-to-Japanese, we used *JTF Translation Quality Evaluation Guidelines* (JTF, 2018) based on Multidimensional Quality Metrics (MQM). We chose four systems for the evaluation and asked a professional translator to evaluate the translations for one talk in the blind test set. We followed the error weighting by a previous study (Freitag et al., 2021a) to calculate error scores. Details of the human evaluation are provided in A.1.2 in Appendix.

The results are shown in Table 16, and we can find the error scores positively correlate with BLEU.

2.7 Future Editions

Possible changes to future editions include:

- changing the latency metric in order to support long unsegmented input.
- extending the task to support speech output.
- removing the text-to-text track in order to consolidate tracks.

3 Offline Speech Translation

Offline speech translation, defined in various forms over the years, is one of the speech tasks with the longest tradition at the IWSLT campaign. This year,[4] it focused on the translation of English audio data extracted from TED talks[5] into text in

[1] Only this language pair has more than one text-to-text systems submitted.

[2] http://utrl.ff.cuni.cz/en

[3] http://iwslt.org/2022/simultaneous

[4] http://iwslt.org/2022/offline

[5] http://www.ted.com

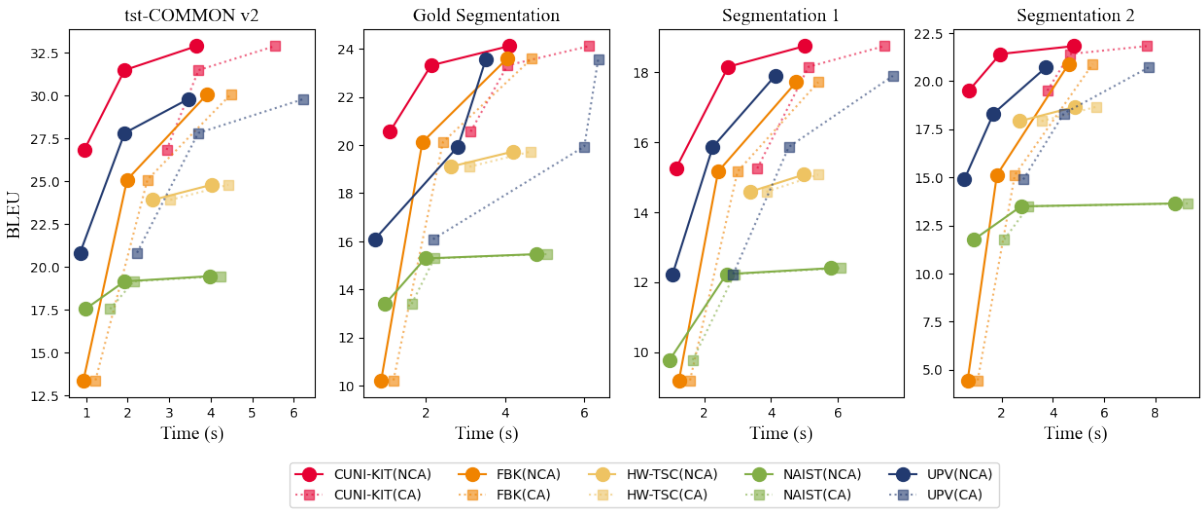

Figure 1: Latency-quality tradeoff curves for English-German.

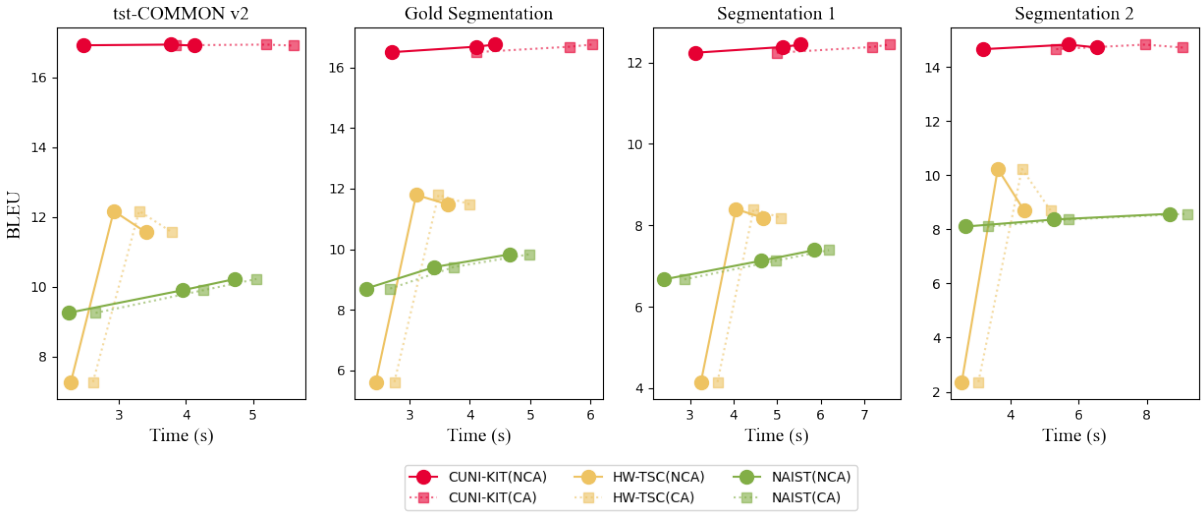

Figure 2: Latency-quality tradeoff curves for English-Japanese.

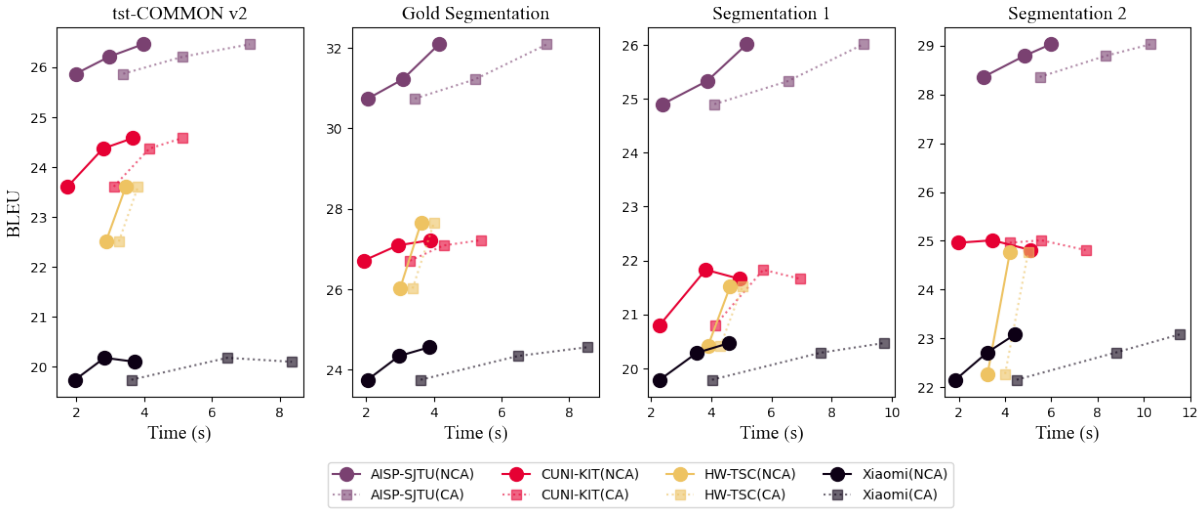

Figure 3: Latency-quality tradeoff curves for English-Mandarin.

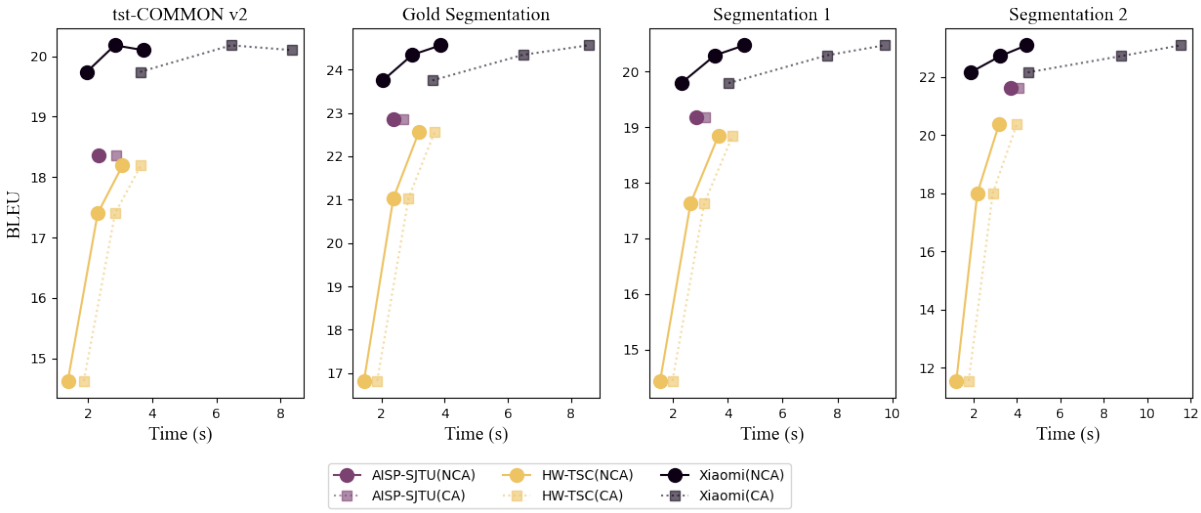

Figure 4: Latency-quality tradeoff curves for English-Mandarin (text-to-text track).

one of the three target languages comprising the 2022 sub-tasks, i.e. German, Japanese, and Mandarin Chinese.

3.1 Challenge

In recent years, offline speech translation (ST) has seen a rapid evolution, characterized by the steady advancement of *direct* end-to-end models (building on a single neural network that directly translates the input audio into target language text) that were able to significantly reduce the performance gap with respect to the traditional *cascade* approach (integrating ASR and MT components in a pipelined architecture). In light of the IWSLT results of the last two years (Ansari et al., 2020; Anastasopoulos et al., 2021) and of the findings of recent work attesting that the gap between the two paradigms has substantially closed (Bentivogli et al., 2021), also this year a key element of the evaluation was to set up a shared framework for their comparison. For this reason, and to reliably measure progress with respect to the past rounds, the general evaluation setting was kept unchanged.

On the architecture side, participation was allowed both with cascade and end-to-end (also known as direct) systems. In the latter case, valid submissions had to be obtained by models that: *i)* do not exploit intermediate symbolic representations (e.g., source language transcription or hypotheses fusion in the target language), and *ii)* rely on parameters that are all jointly trained on the end-to-end task.

On the test set provision side, also this year participants could opt for processing either a precomputed automatic segmentation of the test set or a version of the same test data segmented with their own approach. This option was maintained not only to ease participation (by removing one of the obstacles in audio processing) but also to gain further insights into the importance of properly segmenting the input speech. As shown by the results of recent IWSLT campaigns, effective preprocessing to reduce the mismatch between the provided training material (often "clean" corpora split into sentence-like segments) and the supplied unsegmented test data is in fact a common trait of top-performing systems.

Concerning the types of submission, also this year two conditions were offered to participants: constrained, in which only a pre-defined list of resources is allowed, and unconstrained.

Multiple submissions were allowed, but participants had to explicitly indicate their "primary" (one at most) and "contrastive" runs, together with the corresponding type of system (cascade/end-to-end), training data condition (constrained/unconstrained), and test set segmentation (own/given).

Novelties of the 2022 offline ST task. Within this consolidated overall setting, the organization of this year's task took into consideration new emerging challenges, namely: *i)* the availability of new data covering more language directions, *ii)* the development of new and gigantic pre-trained models, and *iii)* the need for more accurate evaluations. Accordingly, three main differences with

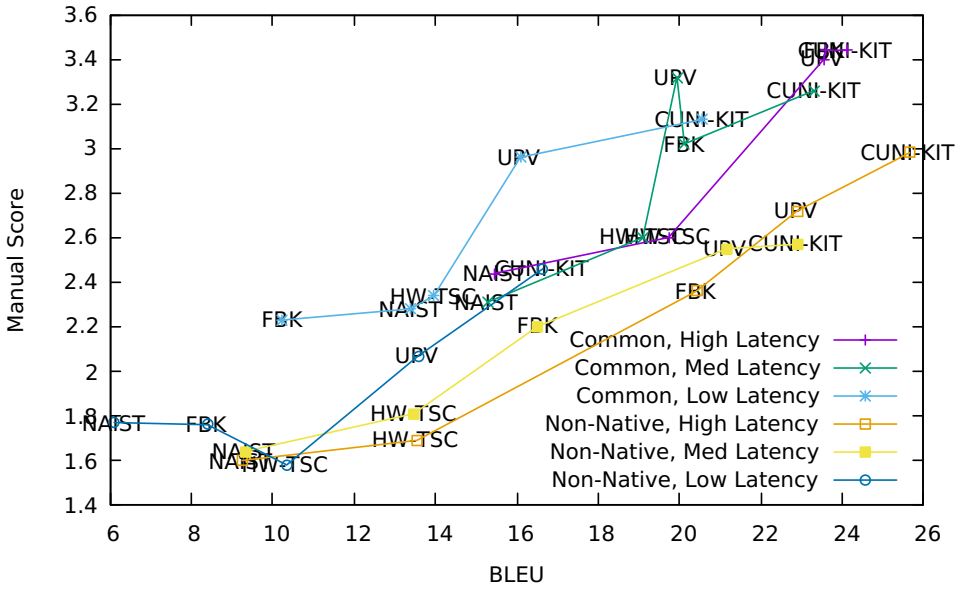

Figure 5: Relation between automatic and manual scoring for English-to-German simultaneous translation on the Common and Non-native part of the test set in the three latency regimes.

respect to previous editions characterize this year's edition:

- To measure systems performance in **different language settings**, two new target languages have been added, extending the number of offline ST sub-tasks to three: English-German (the traditional one), English-Chinese, and English-Japanese.
- To understand the effect of exploiting popular **pre-trained models** in state-of-the-art ST systems, participants were given the possibility to exploit some of them in addition to the allowed training resources for the constrained condition.
- To shed light on the reliability of system ranking based on automatic metrics, and to align our task with other evaluation campaigns (e.g. WMT[6]), the outputs of all the submitted primary systems have been manually evaluated by professional translators. On this basis, a new ranking based on **direct human assessments** was also produced.

3.2 Data and Metrics

Training and development data. Also this year, participants had the possibility to train their systems using several resources available for ST, ASR and MT.

To extend the language directions covered by the offline task, new data was selected from the English-Chinese and English Japanese sections of the MuST-C V2 corpus[7]. For both languages, they include training, dev, and test (Test Common), in the same structure of the MuST-C V2 English-German section (Cattoni et al., 2021) used last year.

Besides the two new language directions of MuST-C V2, also this year the allowed training corpora include:

- MuST-C V1 (Di Gangi et al., 2019);
- CoVoST (Wang et al., 2020a);
- WIT^3 (Cettolo et al., 2012) ;
- Speech-Translation TED corpus[8];
- How2 (Sanabria et al., 2018)[9];
- LibriVoxDeEn (Beilharz and Sun, 2019)[10];
- Europarl-ST (Iranzo-Sánchez et al., 2020);
- TED LIUM v2 (Rousseau et al., 2014) and v3 (Hernandez et al., 2018);

[6] http://www.statmt.org/wmt22/

[7] http://ict.fbk.eu/must-c/

[8] http://i13pc106.ira.uka.de/~mmueller/iwslt-corpus.zip

[9] only English - Portuguese

[10] only German - English

- WMT 2019[11] and 2020[12];
- OpenSubtitles 2018 (Lison et al., 2018);
- Augmented LibriSpeech (Kocabiyikoglu et al., 2018)[13]
- Mozilla Common Voice[14] ;
- LibriSpeech ASR corpus (Panayotov et al., 2015);
- VoxPopuli[15] (Wang et al., 2021).

The only addition over last year is the VoxPopuli dataset.

Similarly to the training data, participants were also provided with a list of pre-trained models that can be used in the constrained condition. The list includes:

- Wav2vec 2.0[16] (Baevski et al., 2020a);
- Hubert[17];
- MBART[18] (Liu et al., 2020);
- MBART50[19] (Tang et al., 2020);
- M2M100[20] (Fan et al., 2021);
- Delta LM[21] (Ma et al., 2021);
- T5[22] (Raffel et al., 2020).

The development data allowed under the constrained condition consist of the dev set from IWSLT 2010, as well as the test sets used for the 2010, 2013, 2014, 2015, 2018, 2019, and 2020 IWSLT campaigns. Using other training/development resources was allowed but, in this case, participants were asked to mark their submission as unconstrained.

Test data. For each language direction, namely En-De, En-Zh and En-Ja, a new test set was created. The new test sets were built from 17 TED talks for En-De, 16 for En-Zh and 13 for En-Ja. None of these talks is included in the current public release of MuST-C. Similar to last year, participants were presented with the option of processing either an unsegmented version (to be split with their preferred segmentation method) or an automatically segmented version of the audio data. For the segmented version, the resulting number of segments is 2,059 (corresponding to about 3h34m of translated speech from 17 talks) for En-De, 1,874 (3h17m) for En-Zh and 1,758 (2h38m) for En-Ja. The details of the three test sets are reported in Table 2.

Lang	Talks	Sentences	Duration
En-De	17	2,059	3h34m
En-Zh	16	1,874	3h17m
En-Ja	13	1,768	2h38m

Table 2: Statistics of the official test sets for the offline speech translation task (*tst2022*).

To measure technology progress with respect to last year's round, participants were asked to process also the undisclosed 2021 En-De test set that, in the segmented version, consists of 2,037 segments (corresponding to about 4.1 hours of translated speech from 17 talks).

Metrics. The systems' performance was evaluated with respect to their capability to produce translations similar to the target-language references. This similarity is measured using the BLEU metric, computed with SacreBLEU (Post, 2018) with default settings.

Similar to the 2021 edition, we consider two different types of target-language references, namely:

- The original TED translations. Since these references come in the form of subtitles, they are subject to compression and omissions to adhere to the TED subtitling guidelines.[23]

[11] http://www.statmt.org/wmt19/
[12] http://www.statmt.org/wmt20/
[13] only English - French
[14] http://voice.mozilla.org/en/datasets – English version en_1488h_2019-12-10
[15] https://github.com/facebookresearch/voxpopuli
[16] https://github.com/pytorch/fairseq/blob/main/examples/wav2vec/README.md
[17] https://github.com/pytorch/fairseq/tree/main/examples/hubert
[18] https://github.com/pytorch/fairseq/blob/main/examples/mbart/README.md
[19] https://github.com/pytorch/fairseq/tree/main/examples/multilingual#mbart50-models
[20] https://github.com/pytorch/fairseq/tree/main/examples/m2m_100
[21] https://github.com/microsoft/unilm/tree/master/deltalm
[22] https://github.com/google-research/text-to-text-transfer-transformer
[23] http://www.ted.com/participate/translate/subtitling-tips

This makes them less literal compared to standard, unconstrained translations;

- Unconstrained translations. These references were created from scratch[24] by adhering to the usual translation guidelines. They are hence exact translations (i.e. literal and with proper punctuation).

Lang Pair	Lang	Sentences	Words
En-De	En	2,059	39,814
	De - Orig	2,059	32,361
	De - Uncon.	2,059	36,655
En-Zh	En	1,874	36,736
	Zh - Orig	1,874	63,876*
	Zh - Uncon.	1,874	64,767*
En-Ja	En	1,768	30,326
	Ja - Orig	1,768	62,778*
	Ja - Uncon.	1,768	74,637*

Table 3: Statistics of the official test set for the offline speech translation task (*tst2022*). * statistics are reported in terms of characters for Chinese and Japanese.

As shown in Table 3, the different approaches to generate the human translations led to significantly different references. For En-De, while the unconstrained translation has a similar length (counted in words) compared to the corresponding source sentence, the original is ~15% shorter in order to fulfil the additional constraints for subtitling. For En-Ja and En-Zh, it is difficult to make a proper comparison with the source data as the Japanese and Chinese data are counted in characters while the English one is counted in words. However, it is evident that the unconstrained translations have more characters than the original ones following a similar trend seen for En-De.

Besides considering separate scores for the two types of references, results were also computed by considering both of them in a multi-reference setting. Similar to last year, the submitted runs were ranked based on case-sensitive BLEU calculated on the test set by using automatic re-segmentation of the hypotheses based on the reference translations by mwerSegmenter.[25]

[24]We would like to thank Meta for providing us with this new set of references.

[25]http://www-i6.informatik.rwth-aachen.de/web/Software/mwerSegmenter.tar.gz

3.3 Submissions

Overall, 10 different teams submitted at total of 29 primary submissions. For the English-to-German task 8 teams submitted 10 runs, for English-to-Chinese 9 teams 11 runs and for the English-to-Japanese task 6 teams participated with 8 primary runs. For all the language pairs two teams submitted a primary cascaded and a primary end-to-end system. Overall, most teams participated in all 3 language directions, partly with individual systems and partly with multi-lingual systems.

We encouraged the submission of end-to-end as well as cascaded systems. Several participants experimented with both types of architectures and in two instances primary end-to-end and cascaded systems were submitted. In total, we had 4 cascaded and 6 end-to-end submissions for the English-to-German tasks, 5 cascaded and 6 end-to-end for English-to-Chinese and 3 cascaded and 5 end-to-end submissions for English-to-Japanese.

One additional change in this year's evaluation campaign was that the use of a list of pre-trained models. Most of the teams investigated this research direction and integrated pre-trained models into their final submission. Both, the integration of pre-trained speech models as well as text models were successfully investigated. In addition, several teams focused on audio segmentation approaches.

- HW-TSC (Li et al., 2022a) submission is built in the cascaded form, including three types of ASR models and one type of translation model. Before performing the speech translation, the LIUM SpkDiarization tool (Rouvier et al., 2013), provided to the participants, was used to cut off the test set wav files into segments. For the ASR part, they use conformer, U2T-transformer and U2-conformer, and all of them are trained on a combination of the MUST-C, COVOST, LibriSpeech, TedLIUM datasets. The system is adapted to the TED domain using domain tags. For the translation model, they trained a Transformer-large on the WMT21-news dataset, and fine-tuned it on the MUST-C and IWSLT datasets. The output of the different ASR models has been re-ranked and the best combination selected as primary submission.

- FBK (Gaido et al., 2022) focused in their

submission on reducing model training costs without sacrificing translation quality. They submitted an end-to-end speech translation system model using the conformer-architecture without pre-trained models. The model is trained on specifically filtered and resegmented parts of the corpus. The final submission is an ensemble of several models.

- USTC-NELSLIP (Zhang et al., 2022b) submitted primary end-to-end and cascaded systems for all three language directions which ensemble several individual models. In the cascaded condition, the ASR models combined transformer and conformer architectures and the MT models are trained on synthetic data to be robust against ASR errors. The end-to-end models also combine conformer and transformer encoders and are partly initialized from ASR systems.

- ALEXA AI (Shanbhogue et al., 2022) submitted an end-to-end speech translation system that leverages pretrained models and cross modality transfer learning for all three language directions. They used encoders for text as well as speech and initialized the models using pretrained speech and text models. The work mainly focused on improving knowledge transfer. In addition, a special focus was put on segmentation strategies.

- NIUTRANS (Zhang et al., 2022c) submission to the English-Chinese track is an end-to-end speech translation system composed of different pre-trained acoustic models and machine translation models. The models were combined by two kinds of adapters and the final submission is an ensemble of three individual speech translation models.

- UPC (Tsiamas et al., 2022a) submission is an end-to-end speech translation model which combines pre-trained speech encoder and text decoder for all the three language directions of the task. As a speech encoder wav2vec 2.0 and HuBERT are used, both already fine-tuned on English ASR data. As a text decoder an mBART50 fine-tuned on multilingual MT (one-to-many) is used. These two modules are coupled with a length adaptor block and in the end-to-end training, additional adapters are trained. For the final submission several initial models are combined.

- KIT (Pham et al., 2022) submitted an end-to-end system using pre-trained audio and text models to all the three language directions. The systems were trained on the initial training data as well as on additional synthetic data. Furthermore, sentence segmentation strategies were investigated. The final submission is an ensemble of several models.

- YI (Zhang and Ao, 2022)) submitted primary end-to-end and cascaded systems for all three language directions using large-scale pre-trained models. Starting from pre-trained speech and language models, the authors investigated a multi-stage pre-training and the use of a task dependent fine-tuning for ASR, MT and speech translation. In addition, various efforts to perform data preparation was carried out. Finally, an ensemble of several models was submitted as the primary submission.

- NEURAL.AI submitted a cascaded speech translation system to the English-to-Chinese speech translation task. The ASR system consists of a conformer encoder and a transformer decoder. The MT system is a fined-tuned deltalm-base.

3.4 Results

This year, the submissions to the IWSLT Offline translation task were not only evaluated using automatic metrics, but also a human evaluation was carried out. All results are shown in detail in the appendix.

3.4.1 Automatic Evaluation

The results for each of the language pairs are shown in the tables in section A.5. For English-to-German we show the results for this year's test set (Table 19) as well as for last year's test set (Table 20). This enables us to also show the progress compared to last year. For the two new language pairs, English-to-Chinese (Table 21) and English-to-Japanese (Table 22), we present the numbers of this year's test set.

First, all the submissions are distributed in a range from 4 to 7 BLEU points. The only exception is Chinese, where one system performed significantly worse than the others. This large

BLEU score range is significantly different than last year's ranking where all the submissions were close to each other. The overall 2022 ranking for the English-German task is quite similar to the ranking obtained for the test set 2021.

Progress The comparison between this year's submissions and last year's submission on test set 2021 in the English-to-German task allows us to measure the progress since last year. As shown in Table 20, 7 out of 9 systems performed better than the best system last year. This year's best system is 4 BLEU points better than last year's system. So, we are seeing a clear improvement in translation quality. One possible reason for the improvement is the additional allowed resources (the Vox-Populi dataset and the pre-trained models). However, also teams not using the additional resources (FBK) outperformed last year's system.

End-to-end vs. cascade As in previous years, we received cascaded and end-to-end submissions. While in the last years, end-to-end systems were able to close the gap to cascaded systems, we do not see this trend since last year. In this year, for all conditions, a cascaded system performed best. Furthermore, when looking at the participants who submitted both, a primary end-to-end and a primary cascaded system, in 6 out of 8 times, the cascaded system performed better than the end-to-end system. Whether this is partly due to the integration of pre-trained models has to be evaluated in further experiments.

Pre-trained models It is difficult to measure the impact of pre-trained models since there is no participant submitting both, a translation system with and without pre-trained models. However, there are some indications of the usefulness of pre-trained models. First, nearly all participants submitted systems with pre-trained models. Typically, these are audio encoders like wav2vec or Hubert for the encoder and text models like mBart for the decoder. Secondly, all winning systems are using this technology. And finally, we see large gains in translation quality compared to last year, where this technique was not allowed. Consequently, these models seem to be an interesting knowledge source. However, it should be noted that the models are rather large and therefore can also be a limiting factor for teams to participate in the evaluation campaign.

Multi-lingual models For the first time, since several years, this year's edition of the offline task included several language directions. Interestingly, this did not lead to a partition of participants into different language pairs, but most participants submitted translations for all three language pairs. While the best performing systems were individually optimized for each language, we also see multilingual models submitted to the tasks. Especially, the integration of pre-trained models, which are typically multi-lingual, made it easier to build translation systems for all three conditions. While the ranking between the languages is not the same, it is still very similar. This indicates that a good system in one language direction typically will also result in good performance in the other directions. While the amount of training resources is at least comparable, this is interesting since the languages are rather different.

3.4.2 Human Evaluation

We conducted a human evaluation of primary submissions based on a random selection of 1,350 segments from the test set of each language pair. Human graders were asked for a direct assessment, expressed through scores between 0 and 100. To minimize the impact of errors in the automatic segmentation, graders were also shown system output for the previous and the following sentence and asked not to let segmentation issues influence their scores. We used Appraise to compute system scores, statistical significance, and rankings. Details of the human evaluation are provided in Section A.2.

As for the results (Tables 23, 24, 25), the ranking of systems matches that of the automatic evaluation when accounting for statistical significance for English to German and English to Chinese, but not for English to Japanese. The scores indicate clear differences between systems (that usually persist across language pairs), but also significant overlap in the translation quality of different systems.

3.4.3 Final remarks

By inspecting this year's results, we can make three final observations.

The first is about the relation between the cascade and end-to-end technology. According to the automatic metrics, and in contrast to last year's campaign, cascade systems achieve the best performance in all the language directions. However,

human evaluation does not validate automatic results for En-De and En-Jp, where the best cascade and end-to-end systems are in the same cluster and not statistically different. This outcome further confirms the findings of Bentivogli et al. (2021) for En-De but extends them to one new language pair out of the two addressed (En-Jp and En-Zh). For this reason, more investigation about the two technologies is still needed and will be further carried out in the next editions of this task.

The other observation is about the introduction of human evaluation in our task. While largely confirming the rankings obtained with automatic metrics, it provides the most reliable picture of the real differences between the systems, showing that they are not so evident as they were detected by automatic metrics. Given the importance of human evaluation to accurately assess state-of-the-art technologies, we plan to rely on it also in the next edition of the task.

The last observation is about the noticeable jump in performance on the progress test set compared to last year's systems. All the current systems have been able to outperform the best 2021 system, with gains reaching up to 6 BLEU score points when using multiple references. While it is difficult to ascribe this improvement to a single factor, it is worth to note that the main change in this year's task setting is the availability of pretrained models. We suggest that these models can have an important role in the final translation quality, and we plan to further investigate their usefulness in the next edition.

4 Speech to Speech Translation

Speech-to-speech translation is the task of translating audio input in a language into audio output in a target language. In the offline setting, systems are able to take into account an entire input audio segment in order to translate, similar to a consecutive interpreter. This is in contrast to streaming or simultaneous settings where systems are only exposed to partial input as in simultaneous interpretation. The goal of this task is to foster the development of automatic methods for offline speech-to-speech translation.

4.1 Challenge

Participants built speech-to-speech translation systems from English into German using any possible method, for example with a cascade system (speech recognition + machine translation + speech synthesis or end-to-end speech-to-text translation + speech synthesis) or an end-to-end or direct system.

4.2 Data and Metrics

Data. This task allowed the same training and testing data from the Offline task on English-German speech-to-text translation to more directly compare Offline S2T and S2ST systems. More details are available in §3.2. We note that while the evaluation data between the two tasks was the same, it was not directly parallel, as different sentence-level segmentation was used. For this task, gold sentence segmentation was used. This means that scores are not directly comparable between the two tasks, though we do evaluate a direct comparison for a subset of submissions.

In addition to the Offline task data, the following training data was allowed to help build German TTS and English-German speech-to-speech models:

- **Synthesized MuST-C**: Target speech for the German target text of MuST-C V2 (Cattoni et al., 2021) which was synthesized for this task using a VITS model (Kim et al., 2021) trained on the German portion of CSS10.
- **CSS10**: A single-speaker German TTS dataset (Park and Mulc, 2019)
- **Pretrained German TTS model**: A pretrained German VITS (Kim et al., 2021) TTS model to facilitate cascaded models and dual submission with the Offline task.

We note that several datasets allowed for the Offline task including Common Voice (Ardila et al., 2020) and LibriVoxDeEn (Beilharz and Sun, 2019) also contain multi-speaker German speech and text data, enabling their use for this task as well.

Metrics. While we evaluate with both automatic and human evaluation scores, systems were ranked according to the human evaluation.

Automatic metrics. To automatically evaluate translation quality, the speech output was automatically transcribed with an ASR system (Conneau et al., 2021),[26] and then BLEU (Papineni

[26] wav2vec2-large-xlsr-53-german

et al., 2002) was computed between the generated transcript and the human-produced text reference. Previous work (Salesky et al., 2021) has shown evaluating synthesized speech with ASR and chrF can be more robust than ASR and BLEU, so we additionally score with chrF (Popović, 2015). All scores were computed using SacreBLEU (Post, 2018).

Human evaluation. Output speech translations were evaluated with respect to translation quality and speech quality.

- **Translation quality:** Bilingual annotators were presented with the source audio and the target audio, and gave scores on the translation quality between 1 and 5. There were 3 annotators per sample and we retained the median score.
- **Output speech quality:** In addition to translation quality (capturing meaning), the quality of the speech output was also human-evaluated along three dimensions: naturalness (voice and pronunciation), clarity of speech (understandability), and sound quality (noise and other artifacts). These axes are more fine-grained than the traditional overall MOS score.

The detailed guidelines for output speech quality were as follows:

- **Naturalness:** Recordings that sound human-like, with natural-sounding pauses, stress, and intonation, should be given a high score. Recordings that sound robotic, flat, or otherwise unnatural should be given a low score.
- **Clarity of speech:** Recordings with clear speech and no mumbling and unclear phrases should be given a high score. Recordings with a large amount of mumbling and unclear phrases should be given a low score.
- **Sound quality:** Recordings with clean audio and no noise and static in the background should be given a high score. Recordings with a large amount of noise and static in the background should be given a low score.

4.3 Submissions

We received submissions from four teams, one of which was withdrawn due to submission errors. We also compare two submissions to the Offline task which were retranslated with the gold segmentation and synthesized using the TTS model provided by the organizers.

MLLP-VRAIN (Iranzo-Sánchez et al., 2022) submitted a cascaded system of separate ASR, MT, and TTS models. They use the same ASR and MT models developed for the Simultaneous ST task, with a less restrictive pruning setup to allow a wider search space for the ASR model and without the multi-path wait-k policy used there for MT. They include a speaker-adaptive module in their TTS system to produce a high quality voice that mimics voice characteristics of the source speaker. Their TTS model is a typical two-stage approach, combining a Conformer-based model (Gulati et al., 2020) to produce spectrograms with a multi-band UnivNet (Jang et al., 2021) model to then produce speech waveforms. They include a speaker encoder, a modified ResNet-34 residual network architecture (He et al., 2016) from (Chung et al., 2018) more widely used for speaker recognition tasks and trained on the TED-LIUM v3 dataset (Hernandez et al., 2018), which is combined with the Conformer output to produce more faithful voices.

HW-TSC (Guo et al., 2022b) submitted a cascaded system of separate ASR, MT, and TTS models. The ASR model ensembles Conformer (Gulati et al., 2020) and S2T-Transformer models (Synnaeve et al., 2020), and is cleaned with the U2 model. The MT model is pretrained on news corpora and finetuned to MuST-C and IWSLT data, with context-aware MT reranking inspired by Yu et al. (2020). They use the provided pretrained VITS TTS model. They use domain tags for each training data source to improve performance. They submitted one primary and three contrastive systems, which ablate individual components. Contrastive1 includes the ASR ensemble but removes reranking for both ASR and MT. Constrastive2 uses the Conformer ASR model only without reranking. Contrastive3 uses the S2T-Transformer ASR model only without reranking.

UPC (Tsiamas et al., 2022a) submitted a cascaded system, extending their direct speech-to-text model submitted to the Offline task with the provided German VITS TTS model for S2ST. Their final speech-to-text model combined initialization using HuBERT models, LayerNorm

and Attention finetuning (LNA), and knowledge distillation from mBART. For both tasks, they used SHAS segmentation during training (Tsiamas et al., 2022b) for consistent improvements. Data filtering and augmentation were also key aspects of their submission.

A direct S2ST model built upon the VITS synthesis model was submitted but withdrawn due to errors.

4.4 Results

Results as scored by automatic metrics are shown in Table 26 and human evaluation results are shown in Table 27 and Table 28 in the Appendix.

Overall results. From the automatic metric perspective, MLLP-VRAIN obtains the highest ASR-BLEU score, followed by HW-TSC and UPC. Note that there is a disagreement between BLEU and chrF ranking for MLLP-VRAIN and HW-TSC. For human evaluation along the speech quality perspective, MLLP-VRAIN obtains a higher quality system compared to the other systems. This is expected as HW-TSC, UPC and the reference system all use the default provided TTS system. It is interesting to note that for these 3 systems, all scores are close to each other on speech quality even though the output content is different. We thus hypothesize that speech quality is orthogonal to translation quality. Finally, for human evaluation along the translation quality perspective, HW-TSC obtained the highest score, followed by MLLP-VRAIN and UPC. Note that this ranking is consistent with the ASR-chrF but not with ASR-BLEU. Surprisingly, the reference system obtains the lowest score. We hypothesize that this may be due to misalignments in the test set between the source audio and the source transcript (rather than between the source transcript and the target translation since the target translations were generated by human translator given the source text transcripts). In addition, we found variance between raters, which could account for this. We will go through a review process for those instances prior to releasing the human judgments.

S2ST Approaches. This year, all systems except the withdrawn submission were cascaded systems, with two systems adopting an ASR + MT + TTS approach and one system adopting an end-to-end S2T + TTS approach. This does not allow us to draw meaningful conclusions on various approaches to the task and we will encourage more direct and/or end-to-end submissions in future editions.

Automatic scoring. To compute automatic metrics, we apply several steps, which may affect quality assessment. The final row of Table 26 shows chrF and BLEU computed on normalized text translations and references; normalizing system output and references reduces scores slightly, by 0.8 BLEU and 0.3 chrF. The larger potential for degradation comes from the synthesis (TTS) and transcription (ASR) roundtrip, which we can directly evaluate the effects of using the reference translations and cascaded systems. Synthesizing the gold reference translation and transcribing with the wav2vec2-large-xlsr-53-german ASR model gives a BLEU score of 68.46 and chrF of 88.78 – degradation of 31.5 BLEU and 11.2 chrF. This confirms errors are introduced by imperfect TTS and ASR models when scoring S2ST systems in this way, and also shows the greater impact of slight variations introduced by TTS and ASR on word-level BLEU than on chrF, which does not necessarily reflect differences in human evaluation (see results in Section B.3). When synthesizing and transcribing machine translation output, there is also degradation in metric scores compared to directly evaluating the text output, but it is considerably smaller. For example, the FBK Offline submission + TTS scores are reduced by 6 BLEU and 4.6 chrF. We see comparing the FBK, KIT, and UPC submissions here, which were all also submitted to the Offline task as speech-to-text systems and then the translations synthesized with the same TTS model, that though there are degradations in performance from synthesis, the relative performance of these models is partly maintained. While the submissions from KIT and FBK both outperform UPC, the relative performance between KIT and FBK reverses according to BLEU – but not according to chrF. This suggests that a finer granularity translation metric may better reflect translation quality after synthesis.

4.5 Conclusion

This is the first time that speech output is introduced in one of the IWSLT shared tasks. The speech-to-speech task serves as a pilot for this kind of task and we plan to run future editions of this task. Possible future extensions include extending

the task to the simultaneous setting and running human evaluations dedicated to additional aspects of the speech output (e.g. preservation of some non-lexical aspects of the input).

5 Low-Resource Speech Translation

This shared task focuses on the problem of developing speech transcription and translation tools for under-resourced languages. For the vast majority of the world's languages there exist little speech-translation parallel data at the scale needed to train speech translation models. Instead, in a real-world situation one might have access to limited, disparate resources (e.g. word-level translations, speech recognition, small parallel text data, monolingual text, raw audio, etc).

Building on last year's task that focused on two varieties of Swahili (Anastasopoulos et al., 2021), the shared task invited participants to build speech translation systems for translating out of two predominantly oral languages, Tamasheq and Tunisian Arabic, and into the linguae francae of the respective regions (English and French). The use of any pre-trained machine translation, speech recognition, speech synthesis, or speech translation model was allowed, as did unconstrained submissions potentially using data other than the ones the organizers provided.

5.1 Data and Metrics

Two datasets were shared for this year's low-resource speech translation track: the Tamasheq-French translation corpus (Boito et al., 2022a), and the Tunisian Arabic-English dataset from the Dialect Translation track (unconstrained condition). In this section we will focus on the Tamasheq corpus, leaving the results for Tunisian Arabic to be presented in Section 6.

The Tamasheq-French translation corpus[27] contains 17 h of speech in the Tamasheq language, which corresponds to 5,829 utterances translated to French. Additional audio data was also made available through the *Niger-Mali audio collection*: 224 h in Tamasheq and 417 h in geographically close languages (French from Niger, Fulfulde, Hausa, and Zarma).[28] For all this data, the speech style is radio broadcasting, and the dataset presents no transcription.

For this track, the main evaluation metric was lower-cased BLEU4 computed over the produced French translation.[29] We also shared with participants results for chrF++. Both are computed on `SacreBLEU` (Post, 2018).[30]

5.2 Submissions

For the Tamasheq language, we received submissions from three teams: ON-TRAC, TALTECH and GMU. We now detail their speech translations models.

ON-TRAC: Boito et al. (2022b) submitted primary and contrastive end-to-end ST systems. Their primary submission focuses on the leveraging of intermediate representations produced by a pre-trained wav2vec 2.0 (Baevski et al., 2020b) base model trained on 234 h of Tamasheq audio. Their end-to-end ST system comprises: a partial wav2vec 2.0 module (in which the last 6 encoder layers were removed), a linear layer for down-projecting the output of the wav2vec 2.0 encoder, and a Transformer decoder with 3 heads, 4 layers and dimensionality of 256. Their contrastive model does not consider SSL features: it uses as input 512-dimensional mel filterbank features. This model leverages *approximate* transcriptions in Tamasheq produced by a French phonemic ASR model. These are used to train an end-to-end ST conformer model that jointly optimizes ASR, MT and ST losses. The model is made of 12 conformer layers of dimensionality 1024, and three transformer decoder layers of dimensionality 2048.

TalTech: Their system is an encoder-decoder ST model with a pretrained XLS-R (Babu et al., 2021) as encoder, and a mBART-50 (Tang et al., 2020) as decoder. For the encoder, they used all the 24 layers of the XLS-R 300M model implemented in `fairseq` (Ott et al., 2019), fine-tuning it on the provided unlabeled raw audio files in Tamasheq (224 h) for 5 epochs. For the decoder, they used the last 12 decoding layers available in the mBART-50 pretrained model.[31] The cross attention layers in the decoder were pointed to the XLS-R's hidden state output to mimic the original cross attention mechanism for text-to-text translation.

[27] https://github.com/mzboito/IWSLT2022_Tamasheq_data

[28] https://demo-lia.univ-avignon.fr/studios-tamani-kalangou/

[29] SacreBLEU BLEU4 signature for the low-resource track: `nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.0.0`

[30] SacreBLEU chrF++ signature for the low-resource track: `nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.0.0`

[31] https://huggingface.co/facebook/mbart-large-50-many-to-many-mmt

GMU: Their model uses the `fairseq S2T` extension (Wang et al., 2020b), using the transformer architecture. They first fine-tune the pre-trained XLS-R 300M encoder on French and Arabic ASR, using portions of the Multilingual TEDx dataset, and then train the whole model on the speech translation task using all provided data.

5.3 Results

All results are presented in Table 4. We observe that the dataset is very challenging: the best achieved BLEU is only 5.7 (ON-TRAC). This challenging setting inspired the teams to leverage pre-trained models: all submissions apply pre-trained initialization for reducing the *cold start* in direct ST in low-resource settings.

Detailing these, ON-TRAC submissions included the training of a wav2vec 2.0 model on target data, and the training of a phonetic French ASR. TalTech used massive multilingual off-the-shelf pre-trained models, and GMU pre-trained their speech encoder on French and Arabic. This illustrates the current trend for ST systems of incorporating pre-trained models. It is nonetheless noticeable that, even with the incorporation of powerful representation extractors (wav2vec 2.0, XLS-R, mBART-50), the achieved results are rather low.

This year's best submission (primary, ON-TRAC) leveraged a Tamasheq wav2vec 2.0 model trained on 234 h. In their post-evaluation results, they included a comparison with different larger wav2vec 2.0 models: XLSR-53 (Conneau et al., 2020), LeBenchmark-7K (Evain et al., 2021), and a multilingual wav2vec 2.0 trained on the Niger-Mali audio collection. Their results hint that smaller pre-trained models focused on the target data seemed to perform better in these low-resource settings. This might be due to the existing domain mismatch between pre-training data (from the off-the-shelf models) and the target data.[32]

The second best submission (contrastive, ON-TRAC) illustrates how even approximate transcriptions can attenuate the challenge of the direct ST task. The authors trained a phonetic French ASR model, and used the produced transcriptions as additional supervision for joint ASR, MT and ST optimization. This solution is very attractive for low-resource settings, as off-the-shelf ASR models – and annotated data to train new ones – are largely available for high-resourced languages.

Finally, we find that TalTech submission illustrates how the application of off-the-box pre-trained multilingual models can be challenging. A similar point can be made about the GMU submission, which despite multilingual finetuning failed to produce meaningful outputs for this challenging task.

In summary, this year's submissions focused on the application of large pre-trained models for end-to-end ST in low-resource settings. They illustrated how low-resource ST remains extremely challenging, even when leveraging powerful speech feature extractors (wav2vec 2.0), and massive multilingual decoders (mBART-50). In such settings, we find that the training of self-supervised models on target data, and the production of artificial supervision (approximate phonemic transcriptions) were the most effective approaches for translating 17 h of Tamasheq audio into French text.

6 Dialect Speech Translation

In some communities, two dialects of the same language are used by speakers under different settings. For example, in the Arabic-speaking world, Modern Standard Arabic (MSA) is used as spoken and written language for formal communications (e.g., news broadcasts, official speeches, religion), whereas informal communication is carried out in local dialects such as Egyptian, Moroccan, and Tunisian. This diglossia phenomenon poses unique challenges to speech translation. Often only the "high" dialect for formal communication has sufficient training data for building strong ASR and MT systems; the "low" dialect for informal communication may not even be commonly written. With this shared task (new for 2022), we hope to bring attention the unique challenges of dialects in diglossic scenarios.

6.1 Challenge

The goal of this shared task is to advance dialectal speech translation in diglossic communities. Specifically, we focus on Tunisian-to-English speech translation (ST), with additional ASR and MT resources in Modern Standard Arabic.

The ultimate goal of this shared task is to explore how transfer learning between "high" and "low" dialects can enable speech transla-

[32] It was previously observed that the wav2vec 2.0 performance degrades when applied to audio data of different speech styles (Conneau et al., 2020).

Team	System	Pre-trained Models	BLEU	chrF++
ON-TRAC	primary	wav2vec 2.0 (Tamasheq)	5.7	31.4
	contrastive	ASR (French)	5.0	26.7
TalTech	primary	XLS-R, mBART-50	2.7	24.3
GMU	primary	XLS-R (Arabic, French)	0.5	16.9

Table 4: Summary of results for the Tamasheq-french corpus for the low-resource shared task.

tion in diglossic communities. Diglossia is a common phenomenon in the world. Besides Arabic vs. its dialects, other examples include Mandarin Chinese vs. Cantonese/Shanghainese/Taiwanese/etc., Bahasa Indonesia vs. Javanese/Sundanese/Balinese/etc., Standard German vs. Swiss German, and Katharevousa vs. Demotic Greek. With this shared task, we imagine that techniques from multilingual speech translation and low-resource speech translation will be relevant, and hope that new techniques that specifically exploit the characteristics of diglossia can be explored.

6.2 Data and Metrics

Participants were provided with the following datasets:

- (a) 160 hours of Tunisian conversational speech (8kHz), with manual transcripts
- (b) 200k lines of manual translations of the above Tunisian transcripts into English, making a three-way parallel data (i.e. aligned audio, transcript, translation) that supports end-to-end speech translation models
- (c) 1200 hours of Modern Standard Arabic (MSA) broadcast news with transcripts for ASR, available from MGB-2 (Specifically, MGB-2 contains an estimated 70% MSA, with the rest being a mix of Egyptian, Gulf, Levantine, and North African dialectal Arabic. All of the MGB-2 train data is allowed.)
- Approximately 42,000k lines of bitext in MSA-English for MT from OPUS (specifically: Opensubtitles, UN, QED, TED, GlobalVoices, News-Commentary).

Datasets (a) and (b) are new resources developed by the LDC, and have been manually segmented at the utterance level. This three-way parallel data (Tunisian speech, Tunisian text, English text) enables participants to build end-to-end or cascaded systems that take Tunisian speech as input and generate English text as final output. The main evaluation metric is lower-cased BLEU on the final English translation[33].

Participants can build systems for evaluation in any of these conditions:

- **Basic condition:** train on datasets (a) and (b) only. This uses only Tunisian-English resources; the smaller dataset and simpler setup makes this ideal for participants starting out in speech translation research.
- **Dialect adaptation condition:** train on datasets (a), (b), (c), (d). The challenge is to exploit the large MSA datasets for transfer learning while accounting for lexical, morphological, and syntactic differences between dialects. This condition may be an interesting way to explore how multilingual models work in multi-dialectal conditions.
- **Unconstrained condition:** participants may use public or private resources for English and more Arabic dialects besides Tunisian (e.g., CommonVoice, TEDx, NIST OpenMT, MADAR, GALE). Multilingual models beyond Arabic and English are allowed. This condition is cross-listed with the low-resource shared task.

The data and conditions available to participants are summarized in Table 5. From the LDC-provided dataset LDC2022E01, we create official train/dev/test1 splits for the basic condition[34] and encourage participants to compare results on "test1." The official blind evaluation set LDC2022E02 is referred to as "test2"; it is collected in the same way as LDC2022E01 and utterance segmentation is given.

[33] SacreBLEU signature for dialect speech translation task: `nrefs:1|case:lc|eff:no|tok:13a|smooth:exp|version:2.0.0`

[34] For datasplit and preprocessing details: https://github.com/kevinduh/iwslt22-dialect

Dataset	Speech	Text (#lines)			Use
	(#hours)	Tunisian	MSA	English	
LDC2022E01 train	160	200k	-	200k	Basic condition
LDC2022E01 dev	3	3833	-	3833	Basic condition
LDC2022E01 test1	3	4204	-	4204	Unofficial evaluation
LDC2022E02 test2	3	4288	-	4288	Official evaluation for 2022
MGB2	1100	-	1.1M	-	Dialect adaptation; mostly MSA
OPUS	-	-	42M	42M	Dialect adaptation condition
Any other data	-	-	-	-	Unconstrained condition

Table 5: Datasets for Dialect Shared Task.

6.3 Submissions

We received submissions from three teams (CMU, JHU, ON-TRAC). Each team explored very different architectures and adaptation techniques. We recommend referring to the system descriptions for details; below is just a brief summary of their contributions:

CMU (Yan et al., 2022) focuses on the Multi-Decoder architecture (Dalmia et al., 2021) implemented in ESPnet, which is an end-to-end ST model that decomposes into ASR and MT subnets while maintaining differentiability. Intuitively, hidden states found by beam search from the ASR decoder are fed as input to the ST encoder. New enhancements on this architecture using hierarcharchical speech encoder and joint CTC/Attention ST decoding are introduced, with gains in BLEU.

Additionally, different approaches to integrating end-to-end and cascaded systems are examined in detailed; for example, one approach uses one system to generate N-best candidates, and the other system to help compute minimum Bayes risk. This resulted in the strongest system for this year's shared task.

In terms of dialect adaptation, the CMU team explored (a) using a Tunisian ASR model select similar MGB2 data by cross-entropy, and (b) using MSA-EN MT trained on OPUS to synthetically augment MGB2 with translations.

JHU (Yang et al., 2022) uses a cascaded architecture, where the ASR component is a conformer-based hybrid attention/CTC model implemented in ESPnet and the MT component is a Transformer model implemented in fairseq. ASR pre-training using wave2vec 2 (XLSR-53) is explored for the unconstrained condition. There is also an emphasis on text normalization to reduce variation in the Tunisian transcripts, which resulted in considerable BLEU gains.

In terms of dialect adaptation, the JHU team investigated a novel data augmentation technique for the MT component: First, a EN→MSA MT model is trained on OPUS and applied to decode LDC2022E01 train set (treating English as source input), synthesizing a paired MSA-Tunisian bitext. With this, a MSA→Tunisian MT model is trained and applied on OPUS, synthesizing a large Tunisian-English bitext. This can be then used in a fine-tuning setup with the original LDC2022E01 data.

ON-TRAC (Boito et al., 2022b) compares both end-to-end and cascaded systems. The end-to-end ST system is a conformer model trained with speed pertubation and SpecAugment, implemented in ESPnet. The cascaded system consists of an ASR component implemented in SpeechBrain, and MT component implemented in fairseq (either biLSTM or convolutional model). Specifically, the ASR component is composed of a wav2vec 2 module, followed by a dense hidden layer and a softmax output of 34 character vocabulary. The use of character outputs in the ASR component is unique to ON-TRAC; other teams employ sub-word units (1000 units for CMU, 400-1000 units for JHU).

In terms of dialect adaptation, the ON-TRAC team explored fine-tuning on the ASR component: first, the ASR model is trained on the MGB2 data; then the model is fine-tuned on the LDC2022E01 data, with the wav2vec portion fixed and the final two layers randomly initialized.

6.4 Results

6.4.1 Automatic evaluation

We are interested in two main scientific questions:

1. For speech translation of primarily spoken dialects, is it beneficial to incorporate data from related dialects with larger written resources? If so, what is the best way to incorporate these resources in training?

2. Does the inherent imbalance and heterogeneity of resources in different dialects favor end-to-end or cascaded architectures? Specifically, there are separate MSA datasets (MGB2, OPUS) that correspond to ASR and MT sub-tasks, but no single MSA dataset that corresponds to an end-to-end speech translation task like the Tunisian-English LDC2022E01 dataset.

Table 29 in the Appendix presents the full results on test2 and test1 sets. Table 6 here presents a summary of select systems in terms of the architecture and training data employed. First, we observe that mixing in MSA/English data tends to improve results over the basic condition of using only the Tunisian/English data. For example, CMU's E2 system obtains 20.8 BLEU, a 0.4 improvement over the E1 system; these are both multi-decoder ensembles, the difference being the training data used. Similarly, JHU's dialect adapt primary system outperforms its basic condition counterpart by 1.8 BLEU. While dialect adaptation is promising, some of the system description papers observe a plateauing effect with additional data, so more work may be needed.

Second, the comparison between end-to-end architectures (directing generating English text from Tunisian speech) vs. cascaded ASR+MT architectures (two stage Tunisian speech to text, followed Tunisian text to English text) is more complex. On one hand, the ON-TRAC system description reports stronger results from its cascaded architecture which exploits wav2vec and additional MGB2 data in its ASR component; on the other hand, the current best-performing model on this task is CMU's E2 system (20.8 BLEU on test2), which mixes both end-to-end and cascaded systems in a Minimum Bayes Risk (MBR) framework. We are not able to make a clear verdict regarding the best architecture for this task, but believe the distinction between end-to-end and cascade architecture may become more blurred in the future.

In summary, we conclude that (1) dialectal adaptation is a promising direction that deserves more research, and (2) the decision between end-to-end vs. cascaded architectures most likely will depend on complicated factors, and both should be pursued during development.

6.4.2 Human evaluation

For the text-based human evaluation in this task, we employed the Direct Assessment (DA) with document context and extended with Scalar Quality Metric (SQM). The overview of the DA+SQM is provided in Section A.4. In this section we only highlight adaptations specific to the task and discuss the results. Since the test set consisted of a few long conversations, human evaluation was run on a subset of it: we sampled 92 excerpts including 10 consecutive segments and used them as document context. We also adapted annotator guidelines for this task asking for judging correct meaning preservation more than grammatical inconsistencies that may appear in informal conversations, as presented on Figure 6.

We have collected 13,860 assessment scores for this task, after excluding quality control items (Table 7). The official results of the human evaluation are presented in Table 31. Systems from each participating teams are significantly different from other teams, but none of the systems was able to provide translation quality competing with the human reference. From the post-annotation survey, some translation issues noticed by annotators were mostly related to incorrect translation of terminology terms and colloquial phrases as well as grammatical and fluency inconsistencies. A few annotators mentioned that in some cases the context of 10 consecutive segments was insufficient and having an access to the original video or audio would help them with the assessment decisions. We will take this feedback into account in next editions of the human evaluation.

7 Formality Control for SLT

Machine translation (MT) models typically return one single translation for each input segment. Specific problems can arise for spoken language translation from English into languages that have multiple levels of formality expressed through honorifics or "grammatical register." For example, the sentence 'Are you sure?' can have two possible correct translations in German: 'Sind Sie sicher?' for the formal register and 'Bist du sicher?' for the informal one. Leaving the model

Team / Condition / System	Architecture	Training Data	BLEU	Δ
CMU / basic / E1	Mix	TA/EN	20.4	-
CMU / dialect adapt / E2	Mix	TA/EN + MSA/EN	20.8	0.4
JHU / basic / primary	Cascaded	TA/EN	17.1	-
JHU / dialect adapt / primary	Cascaded	TA/EN + MSA/EN	18.9	1.8
ON-TRAC / basic /primary	End-to-End	TA/EN	12.4	-
ON-TRAC / unconstrained / post-eval	Cascaded	TA/EN + MSA/EN	14.4	2.0

Table 6: Summary of select systems for Dialect Shared Task (BLEU on test2). We highlight the BLEU improvements (Δ) obtained when training with additional MSA/English data compared with just the Tunisian/English (TA/EN) in the basic condition.

Language pair	Sys.	Ass.	Ass./Sys.
Tunisian→English	7	13,860	1,980

Table 7: Amount of human assessments collected in the text-based evaluation for the Dialect Speech Translation Task run in Appraise. Counts after removing documents with quality control items.

Source	*Could you* provide *your* first name please?
Informal	**Könntest du** bitte **deinen** Vornamen angeben?
Formal	**Könnten Sie** bitte **Ihren** Vornamen angeben?
Source	OK, then please *follow* me to your table.
Informal	ではテーブルまで私について来て。
Formal	ではテーブルまで私についてきてください。
Respectful	ではテーブルまで私についていらしてください。

Table 8: Contrastive translations for EN-DE and EN-JA with different formality. Phrases in bold were annotated by professional translators as marking formality. Example reproduced from Nădejde et al. (2022).

to choose between different valid translation options can lead to translations with inconsistent tone that are perceived as inappropriate by users depending on their demographics and cultural backgrounds, in particular for certain use cases (e.g. customer service, business, gaming chat). Most prior research addressing this problem has been tailored to individual languages and proposed custom models trained on data with consistent formality (Viswanathan et al., 2019), or through side constraints to control politeness or formality (Sennrich et al., 2016; Niu et al., 2018; Feely et al., 2019; Schioppa et al., 2021a).

7.1 Challenge

The goal of this task was to advance research on controlling formality for spoken language translation across multiple diverse target languages and domains.[35] How formality distinctions are expressed grammatically and lexically can vary widely by language. In many Indo-European languages (e.g., German, Hindi, Italian, Russian, and Spanish), the formal and informal registers are distinguished by the second person pronouns and/or corresponding verb agreement. In Japanese, distinctions that express polite, respectful, and humble speech can be more extensive, including morphological markings on the main verb, as well as on some nouns and adjectives; specific lexical choices; and longer sentences. For this task we considered two formality levels: formal and informal. For Japanese, where more than two formality levels are possible, informal was mapped to *kudaketa* and formal to *teineigo*. We give examples of these phenomena in Table 8.

The task focused on text-to-text translation of spoken language with a special theme of zero-shot learning in multilingual models. The task covered supervised and zero-shot settings, both with constrained and unconstrained training data requirements. For the supervised setting, participants were provided with a formality-annotated dataset for training and development for four language pairs: English→German, Spanish, Hindi, Japanese. For the zero-shot task, which covered English→Italian, Russian, only targeted test data was provided after system submission period.

As this was the first shared task organized on formality control, one objective was to establish a standard benchmark including: formality-annotated train and test sets, an evaluation metric, pre-trained baseline models and human evaluation guidelines. To encourage further research in this area and improve the task definition, we will release all these resources (including system outputs and human evaluation annotations) under a shared repository.[36]

[35] https://iwslt.org/2022/formality/

[36] https://github.com/amazon-research/contrastive-controlled-mt/tree/main/

7.2 Data and Metrics

7.2.1 Formality-annotated data

For this task, the organizers provided formality-annotated parallel data comprising of source segments paired with two contrastive reference translations, one for each formality level (informal and formal). The dataset (CoCoA-MT), released by Nădejde et al. (2022), includes phrase-level annotations of formality markers in the target segments in order to facilitate evaluation and analysis (shown in **bold** in Table 8). Formality distinctions are expressed by the use of grammatical register or honorific language. The training set provided to participants comprises segments sourced from two domains: Topical-Chat (Gopalakrishnan et al., 2019) and Telephony. For the test set, organizers additionally included segments sourced from a third held-out domain: Call-Center.

Table 9 reports the number of source segments used for training and evaluation and the overlap between the references (informal vs. formal) as measured by BLEU. The lowest overlap is for Japanese and the highest overlap is for Hindi, indicating that the task of controlling formality is more challenging for Japanese than for Hindi.

Setting	Target	#train	#test	overlap
Supervised	DE	400	600	75.1
	ES	400	600	79.0
	HI	400	600	81.1
	JA	1,000	600	74.6
Zero-shot	IT	0	600	78.8
	RU	0	600	-

Table 9: Number of segments in the training and test data, and overlap between the references in the test set as measured by BLEU (informal vs. formal). Table adapted from Nădejde et al. (2022).

7.2.2 Task definition

Participants were allowed to submit systems under the constrained and unconstrained data settings. To train their systems, participants were allowed to use the formality-labeled dataset provided by the organizers as well as the additional resources described below.

Constrained task: Textual MuST-C v1.2 data (Di Gangi et al., 2019) (for EN-DE, EN-ES, EN-IT, EN-RU), data released for the WMT news translation tasks (WMT21[37] for EN-JA; WMT14[38] for EN-HI), multilingual data from the same dataset (e.g. using EN-FR MuST-C data for training EN-ES models). Participants were not allowed to use external auxiliary tools (e.g., morphological analysers) or pre-trained models (e.g., BERT).

Unconstrained task: Pre-trained models (e.g., mBERT, mBART), additional annotations from morphological analysers, data released by the WMT news translation tasks (WMT21 for EN-DE, EN-RU; WMT13[39] for EN-ES; News Commentary v16[40] and Europarl[41] for EN-IT) and ParaCrawl v9.[42] For EN-HI, EN-JA, participants were allowed to use any other publicly available textual datasets such as WikiMatrix[43] and JParaCrawl.[44]

In both settings, no additional manually created formality-labeled data was allowed. For the unconstrained setting, obtaining additional annotations automatically was allowed as long as the code and data would be publicly released.

Evaluation sets Systems were evaluated for overall quality on MuST-C v1.2 test sets (tst-COMMON) (Di Gangi et al., 2019) for EN→DE, ES, IT, RU. For EN→HI, JA, systems were evaluated on WMT newstest2014 and 2020, respectively. Formality control accuracy was evaluated on the CoCoA-MT formality-annotated test set.

Automatic metrics Overall quality was measured by `sacreBLEU` (Post, 2018) and COMET (Rei et al., 2020). Formality control accuracy was measured using the referenced-based corpus-level metric released with the CoCoA-MT dataset. The metric relies on the contrastive reference translations to automatically assign, with high precision, formality labels (formal vs. informal) to each hypothesis. The segment-level labels are then aggregated to compute the corpus level *Matched-Acccuracy* (M-ACC). For further details on and evaluation of the M-ACC automatic

IWSLT2022/

[37] https://www.statmt.org/wmt21/translation-task.html

[38] https://www.statmt.org/wmt14/translation-task.html

[39] https://www.statmt.org/wmt13/translation-task.html

[40] https://data.statmt.org/news-commentary/v16/

[41] https://www.statmt.org/europarl/

[42] https://paracrawl.eu/

[43] https://opus.nlpl.eu/WikiMatrix.ph

[44] http://www.kecl.ntt.co.jp/icl/lirg/jparacrawl/

metric, we refer the reader to the corresponding CoCoA-MT paper (Nădejde et al., 2022).

7.3 Submissions

We received submissions from three teams. We briefly summarize their methodologies below and refer the reader to their system description papers for details.

ALEXA AI (Zhang et al., 2022a) focused on using data augmentation to generate additional formality data and on using post-editing strategies to convert outputs from a generic NMT system into the desired formality level. They participated in the unconstrained supervised setting for EN→HI, JA. The authors made use of the limited amount of formality data released for the shared task to fine-tune mBART to classify segments as formal or informal. The formality classifier was then used to augment the available training data with additional formal/informal examples which they used to fine-tune a generic NMT system. The final system output from this fine-tuned model was then post-edited using a variety of strategies that the authors examine.

For EN→HI, the post-editing strategy was a rule-based approach which turned informal pronouns to formal pronouns. For EN→JA, the authors focused on a rule-based method for conjugating verbs. Finally, the authors addressed expansion of their methods to something language-agnostic and examined a seq2seq model used to transform formal outputs into informal outputs (they assumed that the output from the fine-tuned model was formal already and the seq2seq model was only used to generate informal translations). Generally, the authors found that the rule-based approaches worked better than the seq2seq post-editing model.

UOS (Vincent et al., 2022) focused on using data augmentation to generate additional formality data and on re-ranking translations from a generic NMT system for a given formality level. They trained systems for all four settings: {constrained, unconstrained} × {supervised, zero-shot}. For the supervised settings, they submitted models for EN→DE, ES. For the zero-shot settings, they submitted models for EN→IT, RU.

In order to augment the formality data, the authors fine-tuned a language model which they used to rank sentences from the available parallel corpora (depending on the constrained or unconstrained setting) by their similarity with the released formal and informal data. Most similar sentences were extracted using a relative position difference algorithm. For the zero-shot case, they noted that a smaller subset of sentences were considered formal (or informal) across the supervised sets for EN→DE, ES. They considered these segments to be strongly formal/informal and used this to find pairs in the zero-shot languages.

They fine-tuned their generic NMT system using the augmented and released formality data. At inference time, they used a large beam width k for beam search and generated k-best hypotheses. The resulting set of hypotheses were re-ranked using a relative frequency model trained on the released formality data (or, for the zero-shot case, using the similar sentences extracted earlier).

UMD (Rippeth et al., 2022) proposed training a single multilingual model that can cover all target languages and formality levels, and experimented with both mBART and mT5 as this model. They also worked with different fine-tuning strategies using both the gold labeled data from the shared task and formality-labeled data extracted from the unlabeled parallel data through rule-based methods or through automatic classification. As fine-tuning strategies they compared using pre-trained models with adapted vector-valued interventions proposed by Schioppa et al. (2021a) against bilingual models optimized towards one formality level (formal or informal) by fine-tuning all model parameters. For automatically labeling data, the authors also relied on fine-tuning a pre-trained multilingual model (XLM-R) for binary classification.

7.4 Results

7.4.1 Automatic Evaluation

In Table 10 and Table 11, we report the formality control accuracy scores (M-ACC) defined in §7.2 for the unconstrained and constrained tracks respectively.[45] For the supervised language arcs (i.e. EN→DE, ES, HI, JA) and unconstrained setting, submitted systems were successfully able to control formality. Average scores across formality settings range from 99.4 for EN→HI to 92.9 for EN→JA. EN→JA was the language pair with the

[45] Here, we focus on results for formality accuracy. We additionally report overall machine translation quality on generic test sets in Table 32 in the appendix along with baseline (uncontrolled) model performance on the formality testset.

Language Pair	System	F	I
EN→DE	UMD	99.4	96.5
	UoS	100.0	100.0
EN→ES	UMD	99.5	93.2
	UoS	98.1	100.0
EN→HI	Alexa AI	99.6	99.8
	UMD	99.4	98.7
EN→JA	Alexa AI	88.8	98.8
	UMD	86.3	97.5
EN→IT	UMD	32.8	97.9
	UoS	51.2	98.6
EN→RU	UMD	100.0	1.10
	UoS	99.5	85.8

Table 10: Formality control accuracy (M-ACC) reported for Formal (F) and Informal (I) for the *unconstrained* task. Note that EN→IT, RU are zero-shot settings.

Language Pair	System	F	I
EN→DE	UoS	100.0	88.6
EN→ES	UoS	87.4	98.0
EN→IT	UoS	29.5	92.9
EN→RU	UoS	98.1	15.4

Table 11: Formality control accuracy (M-ACC) reported for Formal (F) and Informal (I) for the *constrained* task. There was only one system submission by UoS for this track. Note that EN→IT, RU are zero-shot settings.

largest gap between formal and informal accuracy, with both submitted systems doing an average of 11.0 points better on informal translations than formal translations. Finally, we observed that the Alexa AI and UoS teams generally performed better on the supervised unconstrained task than UMD, possibly due to the former's use of high-quality parallel training data as opposed to the latter's use of multilingual pre-trained models.

For the supervised and constrained setting, we had one submission from UoS for EN→DE, ES. On average over both formality settings, their systems achieved an accuracy of 94.3 on EN→DE and 92.7 on EN→ES. For EN→DE, performance was significantly better for formal translations vs. informal translations, while the reverse was true for EN→ES.

In the zero-shot (EN→IT, RU) unconstrained setting, results were more mixed. For the two submissions (from the UMD and UoS teams), there was a clear bias toward one formality level: both

Language Pair	System	F	I
EN→JA	Alexa AI	89.3	92.5
	UMD	82.8	82.7
EN→IT	UMD	13.7	78.3
	UoS	6.0	81.0
EN→RU	UMD	77.2	0.7
	UoS	85.0	71.3

Table 12: Human evaluation of the system level formality accuracy (Formal (F) and Informal (I)) for models in the *unconstrained* setting. Note that EN→IT, RU are zero-shot settings.

Language Pair	System	F	I
EN→IT	UoS	0.2	36.3
EN→RU	UoS	85.3	12.7

Table 13: Human evaluation of the system level formality accuracy (Formal (F) and Informal (I)) for models in the *constrained* setting. Note that EN→IT, RU are zero-shot settings.

systems were better at generating informal Italian and formal Russian translations. This likely reflects the inherent bias toward one formality level in the training set. For the zero-shot constrained setting, only the UoS team submitted a system, and results on the two formality levels were similar, with one formality level outperforming the other. In going from the unconstrained to the constrained setting, the UoS system lost an average of 25 points in accuracy for the zero-shot setting, while only losing 6 points in the fully supervised setting.

7.4.2 Human Evaluation

To complement the automatic evaluations, we conducted human evaluations of formality accuracy for a subset of the language pairs and settings. We selected EN→JA for the unconstrained supervised task, since Japanese has more complex morphological differences between formal and informal translations than the other target languages. We selected both EN→IT, RU for the zero-shot tasks (both constrained and unconstrained).

For each system, we selected a random sample of 300 source segments and collected the formal and informal outputs of the source segments. Annotators were asked to evaluate the outputs and assess whether the translation was formal, informal, neutral, or other.[46] We summarize the results of

[46]We refer the reader to Appendix A.5 for detailed evalua-

the human evaluations here, and give full results in Table 34 in the appendix. System-level accuracy was computed as the number of translations matching their desired formality level divided by the total number of outputs for a given formality level. Inter-annotator agreement as measured by the Krippendorff's α coefficient (Hayes and Krippendorff, 2007) was high, with an average α of 0.89.

Results from the human evaluation of EN→JA for the unconstrained supervised setting were in line with those obtained by the automatic metric: the submitted systems were able to control the formality of the output translations with reasonably high accuracy (90.9 for UMD and 82.8 for ALEXA AI on average across formality levels).

Human evaluation results also corroborated the automatic evaluations for zero-shot formality transfer. The results underscore how challenging the task of zero-shot formality transfer is, with submitted systems generally performing significantly better on one formality level than the other: informal for EN→IT and formal for EN→RU. A notable exception is the UOS EN→RU unconstrained system, which achieves a reasonable accuracy for both formal (85.0) and informal (71.3) registers (again mirroring the findings of the automatic evaluation). Additionally, human evaluators labeled more systems as "neutral" or "other" (i.e., neither formal nor informal) in the zero-shot settings than in the supervised settings.

8 Isometric SLT

Isometric translation is the task of generating translations similar in length to the source input (Lakew et al., 2021b). As a new research area in machine translation, this is the first time isometric translation is proposed as a shared task.[47] We considered 3 translations directions (English - German, English-French and English-Spanish) and 2 training conditions: constrained and unconstrained.

8.1 Challenge

Isometric MT targets issues that emerge when MT is applied to downstream applications such as dubbing, subtitling, and translation of documents. In particular, dubbing requires that the duration of the target speech to be the same of the source in order to achieve isochrony (Lakew et al., 2021b); subtitle translation requires the output to fit blocks of pre-defined length (Matusov et al., 2019); and, finally, document translation requires sometimes to control the translation length in order to preserve the original layout.

We define isometric translations as translations whose length (in characters) is within ±10% of the length of the source (Lakew et al., 2021a). Subjective evaluations of automatically dubbed videos show that isometric translations generated better dubs than translations without any length control (Lakew et al., 2021a).

A few works have focused on controlling the output length of neural MT. Lakew et al. (2019) proposed to split the parallel training data based on target to source length ratio and prepend control tokens. Lakew et al. (2019) and Niehues (2020) incorporated length-encoding mechanisms that adapts positional-encoding (Vaswani et al., 2017) to control the length of the output sequence. Post-hoc approaches have been proposed by Saboo and Baumann (2019) and (Lakew et al., 2021a), where MT system generates an N-best list and then each hypothesis is re-ranked based on its length and score. More recently, Schioppa et al. (2021b) proposed to combine embedding representing attributes (such as length and politeness) with the encoder representation, to control for multiple attributes at generation time; whereas Lakew et al. (2021b) applied self-training to let the model incrementally learn how to generate isometric translations from its own output.

In this shared task, we proposed isometric MT of spoken language transcripts from En → De, Fr, Es. These three directions exhibit different target-to-source length ratios in character count. The length-ratios on the MuST-C training set is 1.12 for En→De, 1.11 for En→Fr, and 1.04 for En→Es.

Shared task participants were invited to work under constrained or unconstrained training regimes and to to submit systems for one or multiple translation directions. When submitting their system outputs, participants were asked to score their performance using a script available for the evaluation period.[48] Participant were also asked to release their outputs under a MIT license to allow for a human evaluation and further analyses.

tion guidelines and label definitions.

[47] https://iwslt.org/2022/isometric

[48] https://github.com/amazon-research/isometric-slt/blob/main/scripts/compute_isometric_slt_stat.sh

	En-De		En-Fr		En-Es	
Test set	LR	LC	LR	LC	LR	LC
MuST-C	1.2	33.2%	1.2	35.2%	1.0	53.2%
Blind	1.1	62.0%	1.1	70.5%	1.0	64.0%

Table 14: Target to source sample length ratio (LR), and length compliance (LC) within a ±10% range, with respect to the source in terms of characters counts, for the MuST-C (tst-COMMON) and blind test sets.

8.2 Data and Metrics

8.2.1 Task Definition

We proposed two types of training regimes:

Constrained task allows the participants to use language pair specific parallel data from the Ted Talks MuST-C v1.2 corpus (Di Gangi et al., 2019). This is an in-domain training data setting for evaluation using the MuST-C test set (tst-COMMON).

Unconstrained Task allows the participants to leverage WMT data, or any other parallel or monolingual data in addition to the MuST-C data which is available under Constrained task. Participants are also allowed to use any pre-trained models like mBART (Liu et al., 2020).[49]

8.2.2 Evaluation Sets

We evaluated isometric machine translation on two test sets:

- MuST-C (tst-COMMON): in-domain test ada that is publicly available for participants to optimize their models.
- Blind Test: a test set of 91 dialogues extracted from 3 YouTube videos.[50] Each dialogue is containing 5-17 utterances is segmented into sentences for a total of 200 sentences. During the evaluation period participants had only access to the source sentences (English).[51]

Target to source sample length ratio and length compliance (±10%) for these test sets are shown in Table 14. The blind dataset was manually post-edited for isometric translation condition i.e. the translators were asked to keep the length of the translation possibly within ±10% of the source length. As a result, it shows a lower length ratio and a higher length compliance than tst-COMMON. Length compliance of the blind set is however not 100% because translators did not find a way to generate translations for many source sentences (phrases) within the range.

8.2.3 Evaluation Metrics

Submissions were evaluated on two dimensions – translation quality and length compliance with respect to the source input.

Translation Quality metrics for isometric translation should be robust to length variations in the hypothesis. For this reason we assessed n-gram metrics such as BLEU (Papineni et al., 2002), and recently proposed semantic based metrics like COMET (Rei et al., 2020) and BERTSCore (Zhang et al., 2019). Our analysis shows that BERTScore is more robust to length variations in the hypothesis when compared with BLEU and COMET. The latter two tends to penalize short hypotheses even for cases where the semantics is preserved. As a result, we primarily use BERTScore to assess translation quality.

Length Compliance (LC) is formulated as the % of translations in the test set that meet the ±10% length criterion. That is, if the source length is 50 characters, a length compliant translation is between 45 to 55 characters. We calculate how many translations fall in this bracket and report the percentage over a test set. In this evaluation, LC is applied only for source samples with length above 10 characters.

8.3 Submissions

We have received four submission from APPTEK, HW-TSC, Amazon Prime Video (APV), and NUV teams.[52] Below we briefly present submitted systems, followed by the baseline approaches we considered for the evaluation.

APPTEK (Wilken and Matusov, 2022) participated in the constrained task for En-De pair. They explored various length controlling approaches with data pre-processing, data augmentation, length tokens as indicators, and multi-pass decoding. For data augmentation, forward and backward translations are applied, together with sample length-targeted pre-processing. For modeling, they combine fine-grained length control token on the en-

[49] https://www.statmt.org/wmt20/index.html

[50] https://github.com/amazon-research/isometric-slt/tree/main/dataset

[51] Dialogue level data and references will be released.

[52] APV team had to withdraw the system paper due to which we are unable to provide a citation.

coder/decoder (Lakew et al., 2019) and length encoding modifying positional encoding (Takase and Okazaki, 2019). As a post-hoc step after translation, the primary system applies a system combination (denoted as length ROVER) over multiple translations from 7 different length classes, ranging from "extra short" to "extra long".

HW-TSC (Li et al., 2022b) participated in the constrained and unconstrained tasks for En-De, and constrained tasks for En-Fr and En-Es. Their submission investigated bi-directional training, R-drop (Wu et al., 2021) (a variant of dropout), data augmentation in forward and backward translation setting, and model ensemble to improve translation quality. For length control they prepended length tokens to the encoder (Lakew et al., 2019), added length ratio based positional encoding (Takase and Okazaki, 2019), applied length aware beam (LAB) to generate N-best lists, and explored different re-ranking strategies. The primary system for HW-TSC was a combination of length token, decoding with LAB and re-ranking of different system outputs. It shows the highest LC score with, however, a tradeoff on translation quality w.r.t. BERTScore.

APV leverages human-in-the-loop mechanism to train an isometric translation model. Their approach builds on top of a multi-source transformer that takes a source and an hypothesis (Tebbifakhr et al., 2018) as input. The hypothesis comes from human post-editing effort for style variation such as matching translation length with the source input. Differently from previous work on interactive post-editing, their work proposes the isometric translation attribute as a new dimension in the human-in-the-loop translation modeling. APV team participated in the unconstrained task for En → De, Fr and Es. Their result shows performance gains against the baseline model when utilizing the post-edited reference as addition model input. However, when adding the isometric criterion for the post-editing stage, translation quality degrades with a slight gain in LC.

NUV (Bhatnagar et al., 2022) participated in the unconstrained task for En-Fr. Their approach is to first translate and then paraphrase. Their MT system is a Marian-NMT system pre-trained on OPUS-MT data (Tiedemann et al., 2020) and fine-tuned on MuST-C training data with three tokens for "short", "normal" and "long" translations. Paraphrases are generated by a MT5 (Xue et al., 2020) model fine-tuned on the PAWS-X paraphrasing data set (Yang et al., 2019).

Baselines: based on the task definition two systems are considered as baselines:

- WEAKBASELINE is a standard neural MT model trained in the constrained data setting, without any isometric translation feature.
- STRONGBASELINE is trained in an unconstrained data setting and implements output length control as in Lakew et al. (2021a) by prepending a length token on the input, generating N-best hypotheses, and re-ranking them with a linear combination of model score and length ratio.

8.4 Evaluations

To assess the performance of isometric translation systems, we measure translation quality and length compliance via automatic and subjective metrics.

8.4.1 Automatic Evaluation

As discussed in Sec. 8.2 we leverage BERTScore and LC metrics to measure isometric translation performance. We take primary system run from each submission and the baseline systems for comparison. Scores are computed against the human post-edited reference of the the blind test set. The automatic evaluation results are given in Table 35.

Translation quality in terms of BERTScore shows that STRONGBASELINE is the best performing system for all directions and training conditions. APPTEK's constrained submission for En-De is the only system performing similarly to STRONGBASELINE. For length compliance, HW-TSC-Constrained shows the best result (LC>=96%) for all pairs. However, the high LC score comes at the cost of lower translation quality with BERTScore.

For the En-De direction, the system from APPTEK-Constrained shows the best trade-off between BERTScore and LC, followed by STRONGBASELINE and HW-TSC-Unconstrained. On En-Fr, NUV-Unconstrained has the best translation quality among all submitted systems in terms of BERTScore but with a significant trade-off on length compliance. On En-Es, APV-Unconstrained shows the highest translation quality but again with a significant trade-off on length

compliance. Over all language pairs, STRONGBASELINE stands out when we look at trade-offs between translation quality and length compliance.

8.4.2 Human Evaluation of Machine Translation Quality

For the text-based human evaluation, we employed the Direct Assessment (DA) with document context and extended with Scalar Quality Metric (SQM). The overview of the DA+SQM is provided in Section A.4. In this section we only highlight modifications specific to the task and discuss the results. The original segmentation was preserved when generating annotation tasks for the human evaluation. In contrast to the Dialect Speech Translation Task, annotators were guided to assess both grammar and meaning of the translations, as presented on Figure 7. The total number of assessment scores collected in text-based human evaluation campaigns per language pair is listed in Table 15.

The official results of the human evaluation are presented in Table 36. Reference translations (TRANSLATOR-A) are significantly better than participating systems and baselines across all three language pairs. In En-De APPTEK-Constrained and the STRONGBASELINE are together in a separate cluster outperforming the rest of the systems. This is also reflected in the automatic metric, where the two systems standout with a higher BERTScore than the other systems. In En-Fr task, a single large cluster includes all systems and baselines. This mean none of the systems were significantly better than the other. In En-Es task, APV-Unconstrained outperformed HW-TSC-Constrained and show similar performance with the STRONGBASELINE.

In the post-annotation questionnaire, most frequently mentioned common issues found in the translation outputs by annotators were: *lack of coherence between segments and inter-sentential translation errors, terminology translation errors and grammatical inconsistencies*. Annotators noticed that one source of those issues was splitting source sentences into short utterances, which automatic systems treated and translated as full sentences.

Language pair	Sys.	Ass.	Ass./Sys.
English→German	7	12,996	1,857
English→French	6	11,286	1,881
English→Spanish	5	9,692	1,938

Table 15: Amount of human assessments collected in the text-based evaluation for the Isometric SLT Task run in Appraise. Counts after removing documents with quality control items.

8.5 Isometric SLT Use case

8.5.1 Automatic Dubbing

As noted in Sec. 8.1, Isometric SLT can be useful for Automatic dubbing that requires the dubbed synthetic speech in the target language to fit the duration of the original speech in the source language. In the previous section, DA+SQM evaluation mainly looked at the translation quality. In this section, using the dubbing architecture of (Federico et al., 2020b) we test the downstream dubbing quality of these translations. To adapt the translations for dubbing, we segment them so as to follow the *speech-pause* arrangement of the source audio using prosodic alignment (PA) (Virkar et al., 2021, 2022). Using the output from PA module, we produce the dubbed audio utilizing a commercial grade Text-to-Speech system with fine-grained duration control (Effendi et al., 2022). We then replace the original audio with the dubbed audio to produce the final dubbed video.

8.5.2 Human evaluation

We generate dubbed videos using all MT outputs and (segmented) post-edited references. To reduce cognitive load, each subject is asked to compare only two MT systems at a time. This results in a total of 31 evaluations across the three dubbing directions, i.e., En-De,Fr,Es. Subjects first watch the dubbed video produced using the reference translation and then rate dubbed videos from two MT outputs. We employed subjects native in the target language and asked them to grade each dubbed video on a scale of 0-10 (0 being the worst and 10 being the best). For each MT system, we compute % Wins, i.e., % subjects preference when comparing two MT systems. For example, if we have 100 clips and according to annotators system A performs better than system B on 60 clips and ties with system B for 10 clips, then %Wins is 60% for system A v/s 30% for system B. We do not use the absolute grading to avoid the bias of each subject

towards dubbing content in general.

For our experiments, we selected 60 dialogues from the blind set, to create 15 video clips such that each clip contains 4 continuous dialogues. To achieve statistically significant results, we employed 15 to 20 subjects (depending on the directions) across all the evaluations.

Table 37 shows the results for % Wins for all 31 evaluations. Additionally, in Table 38, we show the ranking of MT systems based on their performance for the dubbing use case. To rank the systems, we use N_{Wins} that defines the number of evaluations for which a system was preferred over some other system. In general, similar to human assessment for MT quality, we found STRONGBASELINE to be the best system for all three languages and WEAKBASELINE to be the worst for French and Spanish.

Unlike MT human evaluation results, we found WEAKBASELINE to be worse compared to HW-TSC-Constrained even for English-German. In a similar manner, we find that compared to the rankings from MT evaluation, HW-TSCsystems are ranked either higher or on par to APV-Unconstrained and NUV-Constrained. To better understand these differences in the ranking, we computed the Smoothness metric (Federico et al., 2020a) that measures TTS speaking rate stability across contiguous sentences (or phrases) and also consider the LC metric. Note that degraded LC implies that we have either too high or too low speaking rates for the dubbed speech, i.e., LC directly impacts speech fluency (Federico et al., 2020a). Table 39 shows these metrics with systems in a similar order as their ranking. We find that WEAKBASELINE, APV-Unconstrained and NUV-Constrained generally have either a much lower Smoothness or a much lower LC compared to the other systems. This results in poor speaking rate control and impacts % Wins resulting in a different ranking from MT evaluation. The main takeaway is that MT evaluations do not show a complete picture for the downstream task of dubbing as we need not only high quality translations but also translations that permit good speaking rate control.

8.6 Conclusion

This was the first time a shared task on Isometric MT was organized where we looked at evaluated systems on MT quality and length compliance as well as on a downstream task of automatic dubbing which requires isometric translations. With this shared task, we released a new benchmark of manually transcribed and translated scripts (with length compliance in mind) to evaluate isometry in translation. In the possible extensions of this shared task, we plan to include original video along with the transcribed script at dialogue level so that participants can leverage the duration in the source audio to fit the translation within a given time stamp.

Acknowledgements

We would like to thank the IWSLT 2022 sponsors and donors Apple, AppTek, AWS, Meta, Microsoft, and Zoom for supporting the human evaluation of the shared tasks and student participants with computing credits. We would like to thank Mary Johnson, Tom Kocmi and Hitokazu Matsushita for their help with conducting parts of the human evaluation and providing useful comments. We are grateful to the many annotators who participated in the human evaluation and provided their feedback. We would like to thank Zhaoheng Ni, Jeff Hwang and the torchaudio team for providing a streaming ASR model for the simultaneous task. We would like to thank Justine Kao and Brian Bui for running the human evaluation for the speech-to-speech task. The creation of the reference interpretations was funded from the EU project H2020-ICT-2018-2-825460 (ELITR). Ondřej Bojar would like to acknowledge the grant 19-26934X (NEUREM3) of the Czech Science Foundation.

References

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.

Yasuhiro Akiba, Marcello Federico, Noriko Kando, Hiromi Nakaiwa, Michael Paul, and Jun'ichi Tsujii. 2004. Overview of the IWSLT04 Evaluation Campaign. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 1–12, Kyoto, Japan.

Antonios Anastasopoulos, Ondřej Bojar, Jacob Bremerman, Roldano Cattoni, Maha Elbayad, Marcello Federico, Xutai Ma, Satoshi Nakamura, Matteo Negri, Jan Niehues, Juan Pino, Elizabeth Salesky, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Alexander Waibel, Changhan Wang, and Matthew Wiesner. 2021. FINDINGS OF THE IWSLT 2021 EVALUATION CAMPAIGN. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 1–29, Bangkok, Thailand (online). Association for Computational Linguistics.

Ebrahim Ansari, Amittai Axelrod, Nguyen Bach, Ondrej Bojar, Roldano Cattoni, Fahim Dalvi, Nadir Durrani, Marcello Federico, Christian Federmann, Jiatao Gu, Fei Huang, Kevin Knight, Xutai Ma, Ajay Nagesh, Matteo Negri, Jan Niehues, Juan Pino, Elizabeth Salesky, Xing Shi, Sebastian Stüker, Marco Turchi, and Changhan Wang. 2020. Findings of the IWSLT 2020 Evaluation Campaign. In *Proceedings of the 17th International Conference on Spoken Language Translation (IWSLT 2020)*, Seattle, USA.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *LREC*.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al. 2021. XLS-R: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020a. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020b. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

BBC. 2019. BBC Subtitle Guidelines. BBC © 2018 Version 1.1.8.

Benjamin Beilharz and Xin Sun. 2019. LibriVoxDeEn - A Corpus for German-to-English Speech Translation and Speech Recognition.

Luisa Bentivogli, Mauro Cettolo, Marco Gaido, Alina Karakanta, Alberto Martinelli, and Marco Turchi Matteo Negri. 2021. Cascade versus Direct Speech Translation: Do the Differences Still Make a Difference? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, Bangkok, Thailand. Association for Computational Linguistics.

Aakash Bhatnagar, Nidhir Bhavsar, Muskaan Singh, and Petr Motlicek. 2022. Hierarchical Multi-task learning framework for Isometric-Speech Language Translation. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT)*.

Marcely Zanon Boito, Fethi Bougares, Florentin Barbier, Souhir Gahbiche, Loïc Barrault, Mickael Rouvier, and Yannick Estéve. 2022a. Speech resources in the tamasheq language. *Language Resources and Evaluation Conference (LREC)*.

Marcely Zanon Boito, John Ortega, Hugo Riguidel, Antoine Laurent, Loïc Barrault, Fethi Bougares, Firas Chaabani, Ha Nguyen, Florentin Barbier, Souhir Gahbiche, and Yannick Estève. 2022b. ON-TRAC Consortium Systems for the IWSLT 2022 Dialect and Low-resource Speech Translation Tasks. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT)*.

Roldano Cattoni, Mattia Antonino Di Gangi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Must-c: A multilingual corpus for end-to-end speech translation. *Computer Speech & Language*, 66:101155.

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, K. Sudoh, K. Yoshino, and Christian Federmann. 2017. Overview of the IWSLT 2017 Evaluation Campaign. In *Proceedings of the 14th International Workshop on Spoken Language Translation (IWSLT 2017)*, pages 2–14, Tokyo, Japan.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT^3: Web Inventory of Transcribed and Translated Talks. In *Proceedings of the Annual Conference of the European Association for Machine Translation (EAMT)*, Trento, Italy.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, Roldano Cattoni, and Marcello Federico. 2015. The IWSLT 2015 Evaluation Campaign. In *Proceedings of the 12th International Workshop on Spoken Language Translation (IWSLT 2015)*, Da Nang, Vietnam.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico. 2013. Report on the 10th IWSLT Evaluation Campaign. In *Proceedings of the Tenth International Workshop on Spoken Language Translation (IWSLT 2013)*, Heidelberg, Germany.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico. 2014. Report on the 11th IWSLT Evaluation Campaign, IWSLT 2014. In *Proceedings of the Eleventh International Workshop on Spoken Language Translation (IWSLT 2014)*, Lake Tahoe, USA.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico. 2016. The IWSLT 2016 Evaluation Campaign. In *Proceedings of the 13th International Workshop on Spoken Language Translation (IWSLT 2016)*, Seattle, USA.

Colin Cherry and George Foster. 2019. Thinking slow about latency evaluation for simultaneous machine translation. *arXiv preprint arXiv:1906.00048*.

Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. 2018. VoxCeleb2: Deep Speaker Recognition. In *Interspeech*, pages 1086–1090.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised Cross-Lingual Representation Learning for Speech Recognition. In *Proc. Interspeech 2021*, pages 2426–2430.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Siddharth Dalmia, Brian Yan, Vikas Raunak, Florian Metze, and Shinji Watanabe. 2021. Searchable hidden intermediates for end-to-end models of decomposable sequence tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1882–1896, Online. Association for Computational Linguistics.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. MuST-C: a Multilingual Speech Translation Corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2012–2017, Minneapolis, Minnesota.

Matthias Eck and Chiori Hori. 2005. Overview of the IWSLT 2005 evaluation campaign. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 1–22, Pittsburgh, PA.

Johanes Effendi, Yogesh Virkar, Roberto Barra-Chicote, and Marcello Federico. 2022. Duration modeling of neural tts for automatic dubbing. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8037–8041.

Solène Evain, Ha Nguyen, Hang Le, Marcely Zanon Boito, Salima Mdhaffar, Sina Alisamir, Ziyi Tong, Natalia Tomashenko, Marco Dinarelli, Titouan Parcollet, et al. 2021. Task agnostic and task specific self-supervised learning from speech with *LeBenchmark*. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

M. Federico, Y. Virkar, R. Enyedi, and R. Barra-Chicote. 2020a. Evaluating and optimizing prosodic alignment for automatic dubbing. In *Proceedings of Interspeech*, page 5.

Marcello Federico, Luisa Bentivogli, Michael Paul, and Sebastian Stüker. 2011. Overview of the IWSLT 2011 Evaluation Campaign. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 11–27, San Francisco, USA.

Marcello Federico, Mauro Cettolo, Luisa Bentivogli, Michael Paul, and Sebastian Stüker. 2012. Overview of the IWSLT 2012 Evaluation Campaign. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 11–27, Hong Kong, HK.

Marcello Federico, Robert Enyedi, Roberto Barra-Chicote, Ritwik Giri, Umut Isik, Arvindh Krishnaswamy, and Hassan Sawaf. 2020b. From Speech-to-Speech Translation to Automatic Dubbing. In *Proc. of IWSLT*, pages 257–264, Online. ACL.

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.

Weston Feely, Eva Hasler, and Adrià de Gispert. 2019. Controlling Japanese honorifics in English-to-Japanese neural machine translation. In *Proceedings of the 6th Workshop on Asian Translation*, pages 45–53, Hong Kong, China. Association for Computational Linguistics.

Cameron Shaw Fordyce. 2007. Overview of the IWSLT 2007 evaluation campaign. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 1–12, Trento, Italy.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021a. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021b. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.

Ryo Fukuda, Yuka Ko, Yasumasa Kano, Kosuke Doi, Hirotaka Tokuyama, Sakriani Sakti, Katsuhito Sudoh, and Satoshi Nakamura. 2022. NAIST Simultaneous Speech-to-Text Translation System for IWSLT 2022. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT)*.

Marco Gaido, Sara Papi, Dennis Fucci, Giuseppe Fiameni, Matteo Negri, and Marco Turchi. 2022. Efficient yet Competitive Speech Translation: FBK@IWSLT2022. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT)*.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-Chat: Towards knowledge-grounded open-domain conversations. In *Proc. Interspeech 2019*, pages 1891–1895.

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.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented transformer for speech recognition. In *Proceedings of Interspeech 2020, 21st Annual Conference of the International Speech Communication Association*, pages 5036—-5040, Shanghai, China.

Bao Guo, Mengge Liu, Wen Zhang, Hexuan Chen, Chang Mu, Xiang Li, Jianwei Cui, Bin Wang, and Yuhang Guo. 2022a. The Xiaomi Text-to-Text Simultaneous Speech Translation System for IWSLT 2022. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT)*.

Jiaxin Guo, Yinglu Li, Minghan Wang, Xiaosong Qiao, Yuxia Wang, Hengchao Shang, Chang Su, Yimeng Chen, Min Zhang, Shimin Tao, Hao Yang, and Ying Qin. 2022b. The HW-TSC's Speech to Speech Translation System for IWSLT 2022 Evaluation. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT)*.

Andrew Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1:77–89.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia A. Tomashenko, and Yannick Estève. 2018. TED-LIUM 3: twice as much data and corpus repartition for experiments on speaker adaptation. *CoRR*, abs/1805.04699.

Oleksii Hrinchuk, Vahid Noroozi, Abhinav Khattar, Anton Peganov, Sandeep Subramanian, Somshubra Majumdar, and Oleksii Kuchaiev. 2022. NVIDIA NeMo Offline Speech Translation Systems for IWSLT 2022. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT)*.

Javier Iranzo-Sánchez, Jorge Civera Saiz, and Alfons Juan. 2021. Stream-level latency evaluation for simultaneous machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 664–670, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Javier Iranzo-Sánchez, Javier Jorge Cano, Alejandro Pérez-González de Martos, Adrián Giménez Pastor, Gonçal Garcés Díaz-Munío, Pau Baquero-Arnal, Joan Albert Silvestre-Cerdá, Jorge Civera Saiz, Albert Sanchis, and Alfons Juan. 2022. MLLP-VRAIN UPV systems for the IWSLT 2022 Simultaneous Speech Translation and Speech-to-Speech Translation tasks. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT)*.

Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerdà, Javier Jorge, Nahuel Roselló, Adrià Giménez, Albert Sanchis, Jorge Civera, and Alfons Juan. 2020. Europarl-st: A multilingual corpus for speech translation of parliamentary debates. In *Proc. of 45th Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2020)*, pages 8229–8233, Barcelona (Spain).

Won Jang, Dan Lim, Jaesam Yoon, Bongwan Kim, and Juntae Kim. 2021. UnivNet: A Neural Vocoder with Multi-Resolution Spectrogram Discriminators for High-Fidelity Waveform Generation. In *Interspeech*, pages 2207–2211.

Dávid Javorský, Dominik Macháček, and Ondřej Bojar. 2022. Comprehension of subtitles from re-translating simultaneous speech translation.

Japan Translation Federation JTF. 2018. JTF Translation Quality Evaluation Guidelines, 1st Edition (in Japanese).

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *ICML*.

Ali Can Kocabiyikoglu, Laurent Besacier, and Olivier Kraif. 2018. Augmenting Librispeech with French Translations: A Multimodal Corpus for Direct Speech Translation Evaluation. In *Proceedings of LREC 2018*, Miyazaki, Japan.

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.

Surafel Lakew, Marcello Federico, Yue Wang, Cuong Hoang, Yogesh Virkar, Roberto Barra-Chicote, and Robert Enyedi. 2021a. Machine translation verbosity control for automatic dubbing. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Surafel M Lakew, Yogesh Virkar, Prashant Mathur, and Marcello Federico. 2021b. Isometric mt: Neural machine translation for automatic dubbing. *arXiv preprint arXiv:2112.08682*.

Surafel Melaku Lakew, Mattia Di Gangi, and Marcello Federico. 2019. Controlling the output length of neural machine translation. In *Proc. IWSLT*.

Yinglu Li, Minghan Wang, Jiaxin Guo, Xiaosong Qiao, Yuxia Wang, Daimeng Wei, Chang Su, Yimeng Chen, Min Zhang, Shimin Tao, Hao Yang, and Ying Qin. 2022a. The HW-TSC's Offline Speech Translation System for IWSLT 2022 Evaluation. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT)*.

Zongyao Li, JiaXin Guo, Daimeng Wei, Hengchao Shang, Minghan Wang, Ting Zhu, Zhanglin Wu, Zhengzhe Yu, Xiaoyu Chen, Lizhi Lei, Hao Yang, and Ying Qin. 2022b. HW-TSC's Participation in the IWSLT 2022 Isometric Spoken Language Translation. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT)*.

Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Dan Liu, Mengge Du, Xiaoxi Li, Ya Li, and Enhong Chen. 2021. Cross attention augmented transducer networks for simultaneous translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 39–55, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, Alexandre Muzio, Saksham Singhal, Hany Hassan Awadalla, Xia Song, and Furu Wei. 2021. DeltaLM: Encoder-decoder pre-training for language generation and translation by augmenting pretrained multilingual encoders. *arXiv*.

Xutai Ma, Mohammad Javad Dousti, Changhan Wang, Jiatao Gu, and Juan Pino. 2020a. SIMULEVAL: An evaluation toolkit for simultaneous translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 144–150, Online. Association for Computational Linguistics.

Xutai Ma, Juan Pino, and Philipp Koehn. 2020b. SimulMT to SimulST: Adapting simultaneous text translation to end-to-end simultaneous speech translation. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 582–587, Suzhou, China. Association for Computational Linguistics.

Dominik Macháček, Jonáš Kratochvíl, Tereza Vojtěchová, and Ondřej Bojar. 2019. A speech test set of practice business presentations with additional relevant texts. In *Statistical Language and Speech Processing*, pages 151–161, Cham, Switzerland. Springer Nature Switzerland AG.

Evgeny Matusov, Patrick Wilken, and Yota Georgakopoulou. 2019. Customizing neural machine translation for subtitling. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 82–93, Florence, Italy. Association for Computational Linguistics.

J. Niehues, R. Cattoni, S. Stüker, M. Negri, M. Turchi, T. Ha, E. Salesky, R. Sanabria, L. Barrault, L. Specia, and M. Federico. 2019. The IWSLT 2019 Evaluation Campaign. In *Proceedings of the 16th International Workshop on Spoken Language Translation (IWSLT 2019)*, Hong Kong, China.

Jan Niehues. 2020. Machine translation with unsupervised length-constraints. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (AMTA 2020)*, pages 21–35.

Jan Niehues, Roldano Cattoni, Sebastian Stüker, Mauro Cettolo, Marco Turchi, and Marcello Federico. 2018. The IWSLT 2018 Evaluation Campaign. In *Proceedings of the 15th International Workshop on Spoken Language Translation (IWSLT 2018)*, pages 2–6, Bruges, Belgium.

Xing Niu, Sudha Rao, and Marine Carpuat. 2018. Multi-task neural models for translating between styles within and across languages. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1008–1021, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Maria Nădejde, Anna Currey, Benjamin Hsu, Xing Niu, Marcello Federico, and Georgiana Dinu. 2022. CoCoA-MT: A dataset and benchmark for Contrastive Controlled MT with application to formality. In *Findings of the Association for Computational Linguistics: NAACL 2022*, Seattle, USA. Association for Computational Linguistics.

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*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

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 on association for computational linguistics*. Association for Computational Linguistics.

Kyubyong Park and Thomas Mulc. 2019. Css10: A collection of single speaker speech datasets for 10 languages. *Interspeech*.

Michael Paul. 2006. Overview of the IWSLT 2006 Evaluation Campaign. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 1–15, Kyoto, Japan.

Michael Paul. 2008. Overview of the IWSLT 2008 Evaluation Campaign. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 1–17, Waikiki, Hawaii.

Michael Paul. 2009. Overview of the IWSLT 2009 Evaluation Campaign. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 1–18, Tokyo, Japan.

Michael Paul, Marcello Federico, and Sebastian Stüker. 2010. Overview of the IWSLT 2010 Evaluation Campaign. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 3–27, Paris, France.

Ngoc-Quan Pham, Tuan Nam Nguyen, Thai-Binh Nguyen, Danni Liu, Carlos Mullov, Jan Niehues, and Alexander Waibel. 2022. Efficient yet Competitive Speech Translation: FBK@IWSLT2022. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT)*.

Peter Polák, Ngoc-Quan Pham, Tuan Nam Nguyen, Danni Liu, Carlos Mullov, Jan Niehues, Ondřej Bojar, and Alexander Waibel. 2022. System for Simultaneous Speech Translation Task at IWSLT 2022. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT)*.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *arXiv preprint arXiv:2009.09025*.

Elijah Rippeth, Sweta Agrawal, and Marine Carpuat. 2022. Controlling Translation Formality Using Pre-trained Multilingual Language Models. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT)*.

Anthony Rousseau, Paul Deléglise, and Yannick Esteve. 2014. Enhancing the ted-lium corpus with selected data for language modeling and more ted talks. In *LREC*.

M. Rouvier, G. Dupuy, P. Gay, E. Khoury, T. Merlin, and S. Meignier. 2013. An Open-source State-of-the-art Toolbox for Broadcast News Diarization. In *Proceedings of the Interspeech*.

Ashutosh Saboo and Timo Baumann. 2019. Integration of Dubbing Constraints into Machine Translation. In *Proc. of WMT*, pages 94–101, Florence, Italy. ACL.

Elizabeth Salesky, Julian Mäder, and Severin Klinger. 2021. Assessing evaluation metrics for speech-to-speech translation. In *ASRU*.

Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: a large-scale dataset for multimodal language understanding. In *Proceedings of the Workshop on Visually Grounded Interaction and Language (ViGIL)*. NeurIPS.

Andrea Schioppa, David Vilar, Artem Sokolov, and Katja Filippova. 2021a. Controlling machine translation for multiple attributes with additive interventions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6676–6696, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Andrea Schioppa, David Vilar, Artem Sokolov, and Katja Filippova. 2021b. Controlling machine translation for multiple attributes with additive interventions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6676–6696.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40, San Diego, California. Association for Computational Linguistics.

Akshaya Vishnu Kudlu Shanbhogue, Ran Xue, Ching-Yun Chang, and Sarah Campbell. 2022. Amazon Alexa AI's System for IWSLT 2022 Offline Speech Translation Shared Task. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT)*.

Gabriel Synnaeve, Qiantong Xu, Jacob Kahn, Edouard Grave, Tatiana Likhomanenko, Vineel Pratap, Anuroop Sriram, Vitaliy Liptchinsky, and Ronan Collobert. 2020. End-to-end asr: from supervised to semi-supervised learning with modern architectures. In *ICML*.

Sho Takase and Naoaki Okazaki. 2019. Positional Encoding to Control Output Sequence Length. *Proc. of NAACL*.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning.

Amirhossein Tebbifakhr, Ruchit Agrawal, Matteo Negri, and Marco Turchi. 2018. Multi-source transformer with combined losses for automatic post editing. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 846–852.

Jörg Tiedemann, Santhosh Thottingal, et al. 2020. Opus-mt–building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*. European Association for Machine Translation.

Ioannis Tsiamas, Gerard I. Gállego, Carlos Escolano, José A. R. Fonollosa, and Marta R. Costa-jussà. 2022a. Pretrained Speech Encoders and Efficient Fine-tuning Methods for Speech Translation: UPC at IWSLT 2022. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT)*.

Ioannis Tsiamas, Gerard I. Gállego, José A. R. Fonollosa, and Marta R. Costa-jussà. 2022b. Shas: Approaching optimal segmentation for end-to-end speech translation.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *Proceedings of NIPS 2017*.

Sebastian Vincent, Loïc Barrault, and Carolina Scarton. 2022. Controlling Formality in Low-Resource NMT with Domain Adaptation and Re-Ranking: SLT-CDT-UoS at IWSLT2022. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT)*.

Yogesh Virkar, Marcello Federico, Robert Enyedi, and Roberto Barra-Chicote. 2021. Improvements to Prosodic Alignment for Automatic Dubbing. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7543–7574. ISSN: 2379-190X.

Yogesh Virkar, Marcello Federico, Robert Enyedi, and Barra-Chicote Roberto. 2022. Prosodic alignment for off-screen automatic dubbing. *arXiv preprint arXiv:2204.02530*.

Aditi Viswanathan, Varden Wang, and Antonina Kononova. 2019. Controlling formality and style of machine translation output using AutoML. In *SIMBig*, volume 1070 of *Communications in Computer and Information Science*, pages 306–313. Springer.

Changhan Wang, Juan Pino, Anne Wu, and Jiatao Gu. 2020a. Covost: A diverse multilingual speech-to-text translation corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4197–4203.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. 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 993–1003, Online. Association for Computational Linguistics.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Pino. 2020b. fairseq s2t: Fast speech-to-text modeling with fairseq. *arXiv preprint arXiv:2010.05171*.

Minghan Wang, Jiaxin GUO, Yinglu Li, Xiaosong Qiao, Yuxia Wang, Zongyao Li, Chang Su, Yimeng Chen, Min Zhang, Shimin Tao amd Hao Yang, and Ying Qin. 2022. The HW-TSC's Simultaneous Speech Translation System for IWSLT 2022 Evaluation. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT)*.

Patrick Wilken and Evgeny Matusov. 2022. AppTek's Submission to the IWSLT 2022 Isometric Spoken Language Translation Task. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT)*.

Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, Tie-Yan Liu, et al. 2021. R-drop: regularized dropout for neural networks. *Advances in Neural Information Processing Systems*, 34.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Brian Yan, Patrick Fernandes, Siddharth Dalmia, Jiatong Shi, Yifan Peng, Dan Berrebbi, Xinyi Wang, Graham Neubig, and Shinji Watanabe. 2022. CMU's IWSLT 2022 Dialect Speech Translation System. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT)*.

Jinyi Yang, Amir Hussein, Matthew Wiesner, and Sanjeev Khudanpur. 2022. JHU IWSLT 2022 Dialect Speech Translation System Description. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT)*.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. Paws-x: A cross-lingual adversarial dataset for paraphrase identification. *arXiv preprint arXiv:1908.11828*.

Lei Yu, Laurent Sartran, Wojciech Stokowiec, Wang Ling, Lingpeng Kong, Phil Blunsom, and Chris Dyer. 2020. Better document-level machine translation with bayes' rule. *Transactions of the Association for Computational Linguistics*, 8(0):346–360.

Daniel Zhang, Jiang Yu, Pragati Verma, Ashwinkumar Ganesan, and Sarah Campbell. 2022a. Improving Machine Translation Formality Control with Weakly-Labelled Data Augmentation and Post Editing Strategies. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT)*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Weitai Zhang, Zhongyi Ye, Haitao Tang, Xiaoxi Li, Xinyuan Zhou, Jing Yang, Jianwei Cui, Dan Liu, Junhua Liu, and Lirong Dai. 2022b. The USTC-NELSLIP Offline Speech Translation Systems for IWSLT 2022. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT)*.

Yuhao Zhang, Canan Huang, Chen Xu, Xiaoqian Liu, Bei Li, Anxiang Ma, Tong Xiao, and Jingbo Zhu. 2022c. The NiuTrans's Submission to the IWSLT22 English-to-Chinese Offline Speech Translation Task. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT)*.

Ziqiang Zhang and Junyi Ao. 2022. The YiTrans Neural Speech Translation Systems for IWSLT 2022 Offline Shared Task. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT)*.

Qinpei Zhu, Renshou Wu, Guangfeng Liu, Xinyu Zhu, Xingyu Chen, Yang Zhou, Qingliang Miao, Rui Wang, and Kai Yu. 2022. The AISP-SJTU Simultaneous Translation System for IWSLT 2022. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT)*.

Appendix A. Human Evaluation

A Human Evaluation

Human evaluation was carried out for the following tasks: *(i)* Simultaneous Speech Translation, *(ii)* Offline speech translation, *(iii)* Speech to speech translation, *(iv)* Dialect speech translation, *(v)* Isometric SLT, and *(vi)* Formality control for SLT.

Different evaluation protocols were adopted, which are described in the following sections.

A.1 Simultaneous Speech Translation Task

Simultaneous Speech Translation Task ran two different types of manual evaluation: "continuous rating" for English-to-German and MQM for English-to-Japanese.

A.1.1 Human Evaluation for the English-to-German Simultaneous Task

Manual evaluation of English-to-German Simultaneous Task uses a variant of "continuous rating" as described by Javorský et al. (2022).

During the evaluation, bilingual annotators were presented with the source audio and subtitles. The subtitles were displayed in two lines below the audio following the guidelines for video subtitling (BBC, 2019). The annotators were asked to score the quality of the live-presented text output while listening to the input sound. Specifically, the instructions explicitly asked to focus on *content preservation*, or roughly the *adequacy*:

- We ask you to provide your assessment using so-called "continuous rating", which *continuously indicates the quality of the text output given the input utterance you hear* in the range from 1 (the worst) to 4 (the best) by clicking the corresponding buttons or pressing the corresponding keys.

- The rate of clicking/pressing depends on you. However, we suggest clicking *each 5-10 seconds* or when your assessment has changed. We encourage you to provide feedback *as often as possible* even if your assessment has *not changed*.

- The quality scale should reflect primarily the meaning preservation (i.e. evaluating primarily the "content" or very approximately the "adequacy") and the grammaticality and other qualitative aspects like punctuation (i.e. the "form" or extremely roughly the "fluency") should be the secondary criterion.

Context-Aware Judgements One important aspect of the evaluation is that the systems are run independently for each input segment while continuous rating is designed for following the whole speech. Our continuous rating can be thus seen a variant of document-level measure, although the context is (on purpose) available only from the history and not from the future.

When preparing the subtitles from system outputs, we concatenate all sentences into one continuous stream of words.

Time Shift for Better Simultaneity To ease the memory overload of the evaluators, we reduced the delay by shifting the subtitles ahead in time. The shift was done differently for the systems and for the interpretation:

- **Systems**: Each translated sentence was shifted such that its first word was emitted immediately as the source sentence audio began. If there were some words from previous sentence that have not been displayed yet, the emission of the words from the next sentence was delayed. These words were displayed right after all the last word of the previous sentence.

- **Interpreting**: Since we did not have the sentence alignment, we shifted the whole interpretation by a constant such that the last word was emitted with the end of the last uttered word in the source speech. This shift constant was chosen empirically.

Two Test Sets: Common and Non-Native There were two test sets used for the human evaluation: the common test set (consisting of the TED talks used in the Offline Speech Translation task and serving also in the automatic evaluation of Simultaneous Translation task); and a non-native test set. The non-native test set was already used in IWSLT Non-Native Translation Task in 2020 and it is described in Ansari et al. (2020) Appendix A.6. Specifically, we used the Antrecorp (Macháček et al., 2019; mock business presentations by high-school students) and the auditing presentations (SAO) parts.

We show the size of the corpus, as well as the amount of annotation collected in Table 17.

Processing of Collected Rankings Once the results are collected, they are processed as follows. We first inspect the timestamps on the ratings, and remove any that are more than 20 seconds greater than the length of the audio. Because of the natural delay (even with the time-shift) and because the collection process is subject to network and computational constraints, there can be ratings that are timestamped greater than the audio length. If the difference is however too high, we judge it to be an annotation error. We also remove any annotated audio where there is fewer than one rating per 20 seconds, since the annotators were instructed to annotate every 5-10 seconds.

Obtaining Final Scores To calculate a score for each system, we average the ratings across each annotated audio, then average across all the annotated audios pertaining to each system-latency combination. This type of averaging renders all input speeches equally important and it is not affected by the speech length.

The results are shown in Table 18. We observe that, overall, the systems do worse on the non-native audios than they do on the common portion of the test set, whereas the human interpreter performs similarly on both portions.

Indeed some of the high latency systems are rated slightly higher (on average) than the human interpreter on the common portion.

There is a clear effect of latency in almost all systems, with the low-latency subtitles generally rated poorer than the high-latency subtitles by our annotators. This effect is strong in some systems (e.g. FBK) but weaker in others (e.g. NAIST).

A.1.2 MQM-based Human Evaluation for English-to-Japanese Simultaneous Task

For the English-to-Japanese Simultaneous Translation Task, we conducted a human evaluation using a variant of Multidimensional Quality Metrics (MQM). MQM has been used in recent MT evaluation studies (Freitag et al., 2021a) and WMT Metrics shared task (Freitag et al., 2021b). For the evaluation of Japanese translations, we used *JTF Translation Quality Evaluation Guidelines* (JTF, 2018), distributed by Japan Translation Federation (JTF). The guidelines are based on MQM but include some modifications in consideration of the property of the Japanese language.

We hired a Japanese-native professional translator as the evaluator. The evaluator checked translation hypotheses along with their source speech transcripts and chose the corresponding error category and severity for each translation hypothesis using a spreadsheet. Here, we asked the evaluator to focus only on *Accuracy* and *Fluency* errors, because other types of errors in Terminology, Style, and Locale convention would not be so serious in the evaluation of simultaneous translation. Finally, we calculated the cumulative error score for each system based on the error weighting presented by (Freitag et al., 2021a), where *Critical* and *Major* errors are not distinguished.

A.2 Direct Assessment for Offline Speech Translation Task

For the Offline Speech Translation Task (Section 3) we conducted a human evaluation campaign featuring the source-based direct assessment (DA) (Graham et al., 2013; Cettolo et al., 2017; Akhbardeh et al., 2021). In this setting, assessments were performed on a continuous scale between 0 and 100.

Annotation Process We collected segment-level annotations based on the automatic segmentation of the test data. Because we did not want issues from the segmentation to influence scores negatively, we provided translators not only with the source sentence and system translation, but also with the system translation of the previous and following segments. Annotators were then instructed as follows:

"Sentence boundary errors are expected and should not be factored in when judging translation quality. This is when the translation appears to be missing or adding extra words but the source was segmented at a different place. To this end, we have included the translations for the previous and next sentences also. If the source and translation are only different because of sentence boundary issues, do not let this affect your scoring judgement." No video or audio context was provided. Segments were shuffled and randomly assigned to annotators to avoid bias related to the presentation order. Annotations were conducted by a trusted vendor, with professional translators fluent in the source language and native in the target language. For English to German, we additionally collected annotations for the references, which received a considerably higher score than the best submitted system as expected (90.8 vs. 88.9).

Computing rankings System rankings are produced from the average DA scores computed from the average human assessment scores without and with standardization according to each individual annotator's mean and standard deviation, similarly to Akhbardeh et al. (2021). Clusters are identified by grouping together those systems which significantly outperform all others in lower ranking clusters, according to Wilcoxon rank-sum test $p < 0.05$. In Tables 23, 24, and 25 – which show the rankings – clusters are 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.

Official rankings and details on the evaluation campaign for the Offline Speech Translation Task are presented in Section 3.

A.3 Speech to speech translation task

Output speech translations were evaluated with respect to translation quality and speech quality.

- **Translation quality:** Bilingual annotators were presented with the source audio and the target audio, and gave scores on the translation quality between 1 and 5.
- **Output speech quality:** In addition to translation quality (capturing meaning), the quality of the speech output was also human-evaluated along three dimensions: naturalness (voice and pronunciation), clarity of speech (understandability), and sound quality (noise and other artifacts). These axes are more fine-grained than the traditional overall MOS score.

The detailed guidelines for output speech quality were as follows:

- **Naturalness:** Recordings that sound human-like, with natural-sounding pauses, stress, and intonation, should be given a high score. Recordings that sound robotic, flat, or otherwise unnatural should be given a low score.
- **Clarity of speech:** Recordings with clear speech and no mumbling and unclear phrases should be given a high score. Recordings with a large amount of mumbling and unclear phrases should be given a low score.
- **Sound quality:** Recordings with clean audio and no noise and static in the background should be given a high score. Recordings with a large amount of noise and static in the background should be given a low score.

A.4 Direct Assessment with Scalar Quality Metric for the Dialect and Isometric Speech Translation Tasks

For the Dialect Speech Translation Task (Section 6) and Isometric SLT Task (Section 8) we piloted a human evaluation campaign featuring the source-based direct assessment (DA) (Graham et al., 2013; Cettolo et al., 2017; Akhbardeh et al., 2021) with document context extended with Scalar Quality Metric (SQM) (Freitag et al., 2021a). In this setting, 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 annotator guidelines based on those proposed by Freitag et al. (2021a). SQM helped standardizing scores across annotators.

Tool We used the Appraise evaluation framework[53] (Federmann, 2018) for collecting segment-level judgements within document context. No video or audio context was provided. Annotation guidelines were adapted specifically for each task as described in Sections 6 and 8. Screenshots of an example annotation for the Dialect and Isometric Speech Translation Tasks are presented on Figures 6 and 7.

Task generation A single task consisted of 100 segments from around 10 documents. Human references were included as additional system output to provide an estimate of human performance. Each individual annotator completed between 4 and 8 tasks. Whenever possible, we assigned tasks to annotators making sure that one annotator evaluates outputs from all systems on the same subset of the test set. This increased repetitiveness, but potentially improved consistency of assessments across systems.

Annotation and quality control All annotators were either professional translators or linguists fluent in the source language and native in the target language or linguists, and the majority of them had previous experience in the evaluation of translation outputs.[54] Although our annotators were professionals, we employed a standard quality filtering procedure. Around 10% of segments in each task were quality control items in the form of bad reference pairs distributed usually across one or two documents. Please refer to (Akhbardeh et al., 2021) for more details on the generation of bad references. Assessments of an annotator who has not demonstrated ability to reliably score degraded translations significantly lower than corresponding original system outputs using a paired significance test with $p < 0.05$ would be omitted from the evaluation. As expected, none of our annotators appeared unreliable.

We have collected 47,834 assessments. This number already excludes documents with quality control items, which provides almost 2,000 annotations per system, including references.

Computing rankings System rankings are produced from the average DA scores computed from the average human assessment scores without and with standardization according to each individual annotator's mean and standard deviation, similarly to Akhbardeh et al. (2021). We exclude entire documents with one or more quality control items from ranking computation. Clusters are identified by grouping those systems together which significantly outperform all others in lower ranking clusters, according to Wilcoxon rank-sum test $p < 0.05$. In Tables 31 and 36 – which show the rankings – clusters are 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.

Official rankings and details on the evaluation campaign for the Dialect Speech Translation Task and Isometric SLT Task are presented respectively in Sections 6 and 8.

A.5 Formality Control

In this section, we reproduce the instructions given to the translators for IT, JA and RU for the formality control shared task. Instructions for JA are similar but include some language-specific notes. For brevity, we also remove example translations show to the translators.

Overview We would like to annotate multiple system outputs. For each of the 300 sentence ids (sid) there are 4-6 system outputs - please shuffle the order of the systems when showing it to annotators. We would like two annotators per target language.

Guidelines You will be shown an English source sentence and a machine translation of the source sentence. Your task will be to label the translation based on the formality level. Note that labels that you generate will be on the sentence level (one label per sentence). For example, given the source sentence "It was nice chatting with you, have a great night!" and a translation "Es war schön, mit Ihnen zu plaudern, haben Sie eine tolle Nacht!", you would label the example based on the formality level of the translation as one of *Formal*, *Informal*, *Neutral*, *Other*.

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

[54] In the post annotation questionnaire, 57% of annotators indicated their experience as high (evaluating MT outputs regularly) and 32% as moderate (did it more than few times).

Special Cases to Consider

1. Only label formality level, and ignore other mistakes such as a wrong sense.
2. Only label based on the formality level of the translation. Note that we don't want to label whether the formality level is correct in translation, but rather which formality level is marked in the translation.
3. If at least one word in the source is not translated at all and some meaning is lost, then label the translation as Other.

Label Categories

1. **Formal** – The formality level is consistently Formal in the translation.
2. **Informal** – The formality level is consistently Informal in the translation.
3. **Neutral** – The translation is phrased in a way that does not explicitly express a formality level.
4. **Other** – Explain the reason in the Notes section.
   - The formality level is inconsistent such as using both formal and informal pronouns.
   - If at least one word in the source is not translated at all and should have been marked in the target language for formality and some meaning is lost.
   - If you feel strongly that the translation does not fit into any of the cases listed above, please label it as "other" and explain the reason in the Notes section.

Appendix B. Evaluation Results and Details

B.1. Simultaneous Speech Translation

Automatic Evaluation Results

· Summary of the results of the simultaneous speech translation for **English-German.**
· Results are reported on the blind test set and systems are grouped by latency regime (set on tst-COMMON v2)
· For each entry for latency metric, the upper one is non computation aware, while the lower one is computation aware.
· BLEU number in parenthesis indicate that the system does not satisfy the latency constraints.
· Raw system logs are also provided on the task web site.[55]

	Low Latency				Medium Latency				High Latency			
Team	BLEU	AL	AP	DAL	BLEU	AL	AP	DAL	BLEU	AL	AP	DAL
tst-COMMON v2												
CUNI-KIT	**26.82**	**0.96** **2.94**	**0.77** **1.52**	**2.07** **6.38**	**31.47**	**1.93** **3.71**	**0.86** **1.39**	**2.96** **5.80**	**32.87**	**3.66** **5.54**	**0.96** **1.37**	**4.45** **6.61**
FBK	13.38	0.94 1.23	0.58 0.66	1.31 1.47	25.08	1.99 2.48	0.80 0.93	2.36 2.79	30.07	3.92 4.49	0.95 1.09	4.15 4.70
HW-TSC	(18.56)	1.96 2.39	0.79 0.92	2.41 2.82	23.90	2.61 3.03	0.87 1.01	3.07 3.49	24.78	4.02 4.42	0.96 1.10	4.31 4.71
NAIST	17.54	0.99 1.58	0.68 0.87	1.50 2.43	19.15	1.93 2.15	0.82 0.91	3.63 3.99	19.45	3.98 4.23	0.94 1.01	5.17 5.50
UPV	20.82	0.86 2.23	0.70 1.18	1.43 3.71	27.80	1.93 3.70	0.83 1.43	2.34 5.06	29.78	3.46 6.23	0.93 1.71	3.71 7.53
Gold Segmentation												
CUNI-KIT	**20.56**	**1.09** **3.13**	**0.76** **1.46**	**2.25** **6.69**	**23.31**	**2.13** **4.06**	**0.85** **1.37**	**3.24** **6.27**	**24.11**	**4.10** **6.12**	**0.96** **1.36**	**4.92** **7.29**
FBK	10.23	0.87 1.18	0.54 0.61	1.28 1.42	20.12	1.91 2.43	0.78 0.89	2.37 2.79	23.59	4.05 4.67	0.95 1.07	4.36 4.93
HW-TSC	(13.97)	1.91 2.39	0.77 0.89	2.47 2.91	19.10	2.62 3.10	0.86 0.99	3.18 3.66	19.73	4.20 4.65	0.95 1.09	4.57 5.00
NAIST	13.40	0.97 1.64	0.67 0.85	1.55 2.60	15.29	1.98 2.21	0.82 0.89	3.96 4.35	15.47	4.80 5.07	0.96 1.02	5.79 6.14
UPV	16.09	0.71 2.18	0.68 1.13	1.42 3.78	19.94	2.81 6.00	0.84 1.58	3.36 7.76	23.55	3.51 6.35	0.92 1.63	3.85 7.82
Segmentation 1												
CUNI-KIT	**15.25**	**1.16** **3.59**	**0.75** **1.47**	**2.67** **7.23**	**18.15**	**2.72** **5.12**	**0.86** **1.36**	**3.98** **6.99**	**18.74**	**5.00** **7.38**	**0.97** **1.37**	**5.67** **8.16**
FBK	9.20	1.25 1.58	0.60 0.66	1.95 2.14	15.16	2.42 3.00	0.80 0.91	3.07 3.58	17.71	4.75 5.41	0.96 1.07	5.08 5.71
HW-TSC	(10.66)	2.65 3.10	0.79 0.88	3.23 3.59	14.58	3.37 3.86	0.87 0.99	3.94 4.36	15.07	4.98 5.40	0.96 1.08	5.32 5.71
NAIST	9.78	0.97 1.66	0.65 0.82	1.75 2.66	12.23	2.67 2.91	0.83 0.89	4.30 4.67	12.40	5.78 6.08	0.98 1.03	6.26 6.59
UPV	12.23	1.06 2.87	0.68 1.14	1.86 4.45	15.86	2.26 4.53	0.80 1.35	2.87 5.91	17.89	4.12 7.64	0.93 1.67	4.51 8.86
Segmentation 2												
CUNI-KIT	**19.51**	**0.73** **3.79**	**0.66** **1.43**	**2.71** **11.29**	**21.41**	**1.95** **4.67**	**0.74** **1.28**	**4.10** **9.69**	**21.82**	**4.81** **7.66**	**0.88** **1.29**	**7.06** **11.31**
FBK	4.45	0.68 1.07	0.34 0.39	1.17 1.30	15.12	1.82 2.52	0.61 0.69	2.65 3.17	20.89	4.62 5.56	0.85 0.96	5.50 6.35
HW-TSC	(12.53)	1.92 2.66	0.63 0.74	2.81 3.58	17.92	2.71 3.56	0.75 0.88	3.77 4.75	18.66	4.86 5.68	0.86 1.00	5.84 6.73
NAIST	11.77	0.93 2.11	0.60 0.83	1.92 4.32	13.49	2.76 3.05	0.84 0.90	7.75 8.42	13.64	8.76 9.26	0.97 1.03	10.62 11.23
UPV	14.89	0.55 2.85	0.62 1.03	1.78 5.84	18.32	1.69 4.43	0.70 1.17	2.71 7.29	20.72	3.74 7.75	0.82 1.48	4.62 11.16

[55] https://iwslt.org/2022/simultaneous

· Summary of the results of the simultaneous speech translation for **English-Japanese.**
· Results are reported on the blind test set and systems are grouped by latency regime (set on tst-COMMON v2)
· For each entry for latency metric, the upper one is non computation aware, while the lower one is computation aware.
· Raw system logs are also provided on the task web site.[56]

	Low Latency				**Medium Latency**				**High Latency**			
Team	BLEU	AL	AP	DAL	BLEU	AL	AP	DAL	BLEU	AL	AP	DAL
tst-COMMON v2												
CUNI-KIT	**16.92**	**2.46** **3.84**	**0.90** **1.38**	**3.22** **5.45**	**16.94**	**3.77** **5.20**	**0.97** **1.34**	**4.29** **6.03**	**16.91**	**4.13** **5.61**	**0.98** **1.34**	**4.53** **6.20**
HW-TSC	7.27	2.28 2.61	0.81 0.92	2.68 2.91	12.17	2.92 3.30	0.92 1.06	3.38 3.71	11.56	3.40 3.79	0.95 1.09	3.84 4.16
NAIST	9.25	2.24 2.65	0.88 1.03	3.04 3.50	9.90	3.95 4.26	0.96 1.07	4.59 4.94	10.22	4.73 5.05	0.99 1.09	4.96 5.30
Gold Segmentation												
CUNI-KIT	**16.50**	**2.71** **4.10**	**0.90** **1.37**	**3.35** **5.79**	**16.68**	**4.10** **5.66**	**0.97** **1.34**	**4.57** **6.48**	**16.75**	**4.42** **6.02**	**0.98** **1.34**	**4.80** **6.67**
HW-TSC	5.62	2.44 2.75	0.79 0.89	2.71 2.92	11.79	3.11 3.48	0.91 1.04	3.46 3.80	11.48	3.63 4.00	0.95 1.08	3.96 4.30
NAIST	8.70	2.28 2.68	0.86 0.99	2.89 3.40	9.41	3.41 3.73	0.94 1.04	4.46 4.87	9.83	4.66 4.98	0.98 1.06	5.08 5.44
Segmentation 1												
CUNI-KIT	**12.24**	**3.12** **4.99**	**0.87** **1.34**	**4.22** **7.14**	**12.38**	**5.12** **7.17**	**0.97** **1.33**	**5.79** **8.10**	**12.44**	**5.54** **7.58**	**0.98** **1.33**	**6.03** **8.22**
HW-TSC	4.15	3.25 3.63	0.79 0.87	3.75 4.01	8.40	4.05 4.46	0.91 1.01	4.55 4.89	8.18	4.68 5.09	0.95 1.05	5.14 5.49
NAIST	6.67	2.40 2.87	0.81 0.92	3.35 3.90	7.13	4.64 4.98	0.93 1.00	5.56 5.97	7.39	5.86 6.19	0.98 1.04	6.23 6.58
Segmentation 2												
CUNI-KIT	**14.65**	**3.19** **5.34**	**0.77** **1.27**	**4.54** **9.80**	**14.82**	**5.71** **7.95**	**0.90** **1.29**	**7.37** **11.45**	**14.71**	**6.55** **9.06**	**0.93** **1.30**	**8.11** **12.03**
HW-TSC	2.36	2.56 3.05	0.52 0.58	2.99 3.26	10.23	3.62 4.33	0.76 0.87	4.38 5.01	8.70	4.39 5.17	0.82 0.94	5.30 5.96
NAIST	8.10	2.67 3.32	0.73 0.85	3.81 4.82	8.36	5.28 5.71	0.91 0.99	9.00 9.72	8.57	8.69 9.20	0.97 1.03	10.32 10.94

[56] https://iwslt.org/2022/simultaneous

· Summary of the results of the simultaneous speech translation for **English-Mandarin.**
· Results are reported on the blind test set and systems are grouped by latency regime (set on tst-COMMON v2)
· For each entry for latency metric, the upper one is non computation aware, while the lower one is computation aware.
· BLEU number in parenthesis indicate that the system does not satisfy the latency constraints.
· Raw system logs are also provided on the task web site.[57]

	Low Latency				**Medium Latency**				**High Latency**			
Team	BLEU	AL	AP	DAL	BLEU	AL	AP	DAL	BLEU	AL	AP	DAL
tst-COMMON v2												
AISP-SJTU	**25.87**	**1.99** **3.39**	**0.87** **1.81**	**3.35** **6.53**	**26.21**	**2.97** **5.14**	**0.94** **1.97**	**4.16** **7.80**	**26.46**	**3.97** **7.12**	**0.98** **2.05**	**4.62** **8.42**
CUNI-KIT	23.61	1.75 3.11	0.85 1.34	2.56 4.77	24.37	2.79 4.16	0.93 1.34	3.49 5.32	24.58	3.67 5.12	0.97 1.34	4.22 5.88
HW-TSC	(18.60)	2.18 2.56	0.84 0.97	2.66 2.93	22.51	2.88 3.26	0.92 1.06	3.33 3.62	23.60	3.46 3.82	0.95 1.09	3.81 4.10
Xiaomi	19.74	1.97 3.63	0.83 1.32	2.64 4.82	20.18	2.84 6.46	0.90 2.18	3.62 9.68	20.10	3.73 8.36	0.95 2.31	4.18 10.81
Gold Segmentation												
AISP-SJTU	**30.74**	**2.05** **3.44**	**0.86** **1.56**	**3.46** **6.72**	**31.22**	**3.08** **5.22**	**0.93** **1.72**	**4.34** **8.06**	**32.09**	**4.15** **7.34**	**0.97** **1.81**	**4.83** **8.75**
CUNI-KIT	26.71	1.92 3.29	0.83 1.32	2.65 5.09	27.09	2.93 4.29	0.92 1.31	3.62 5.57	27.22	3.90 5.39	0.97 1.32	4.44 6.23
HW-TSC	(19.83)	2.25 2.66	0.82 0.95	2.68 2.98	26.02	3.00 3.37	0.91 1.04	3.43 3.72	27.65	3.62 4.00	0.95 1.08	3.97 4.29
Xiaomi	23.75	2.04 3.61	0.82 1.28	2.62 4.78	24.34	2.97 6.48	0.90 2.11	3.71 9.86	24.56	3.87 8.55	0.95 2.28	4.29 11.15
Segmentation 1												
AISP-SJTU	**24.90**	**2.39** **4.11**	**0.83** **1.41**	**4.12** **7.78**	**25.33**	**3.87** **6.56**	**0.93** **1.60**	**5.30** **9.57**	**26.01**	**5.18** **9.04**	**0.97** **1.70**	**5.93** **10.48**
CUNI-KIT	20.80	2.29 4.13	0.81 1.27	3.51 6.30	21.83	3.82 5.73	0.92 1.30	4.79 7.16	21.66	4.95 6.96	0.97 1.31	5.66 7.81
HW-TSC	(16.09)	3.03 3.47	0.82 0.91	3.68 3.99	20.42	3.90 4.31	0.91 1.00	4.50 4.80	21.52	4.63 5.04	0.95 1.05	5.11 5.43
Xiaomi	19.79	2.30 4.03	0.79 1.19	3.20 5.43	20.29	3.53 7.62	0.89 1.97	4.57 11.32	20.47	4.60 9.72	0.94 2.09	5.25 12.54
Segmentation 2												
AISP-SJTU	**28.36**	**3.06** **5.50**	**0.83** **1.50**	**7.10** **14.52**	**28.79**	**4.82** **8.33**	**0.91** **1.64**	**8.71** **16.96**	**29.03**	**5.97** **10.29**	**0.94** **1.70**	**9.26** **17.68**
CUNI-KIT	24.96	1.97 4.20	0.70 1.20	3.41 8.54	25.01	3.46 5.57	0.80 1.21	5.19 9.32	24.81	5.11 7.48	0.88 1.25	7.01 10.79
HW-TSC	(13.80)	2.26 2.93	0.59 0.68	3.00 3.39	22.27	3.24 4.00	0.74 0.85	4.21 4.70	24.77	4.21 5.00	0.82 0.93	5.21 5.76
Xiaomi	22.15	1.85 4.50	0.69 1.19	3.04 8.10	22.71	3.23 8.80	0.77 2.10	4.84 18.63	23.08	4.43 11.55	0.83 2.30	5.63 21.16

[57] https://iwslt.org/2022/simultaneous

· Summary of the results of the simultaneous speech translation for **text-to-text track, English-Mandarin**
· The input of the each system is the output from the provided streaming ASR model, and the latency is evaluated in seconds.
· Results are reported on the blind test set and systems are grouped by latency regime (set on tst-COMMON v2)
· For each entry for latency metric, the upper one is non computation aware, while the lower one is computation aware.
· Raw system logs are also provided on the task web site.[58]

	Low Latency				**Medium Latency**				**High Latency**			
Team	BLEU	AL	AP	DAL	BLEU	AL	AP	DAL	BLEU	AL	AP	DAL
tst-COMMON v2												
AISP-SJTU					18.36	2.35 2.89	0.88 1.05	4.04 4.83				
HW-TSC	14.63	1.38 1.88	0.73 0.86	2.01 2.43	17.40	2.31 2.85	0.86 1.00	2.90 3.37	18.19	3.08 3.65	0.92 1.07	3.57 4.08
Xiaomi	19.74	1.97 3.63	0.83 1.32	2.64 4.82	20.18	2.84 6.46	0.90 2.18	3.62 9.68	20.10	3.73 8.36	0.95 2.31	4.18 10.81
Gold Segmentation												
AISP-SJTU					22.85	2.38 2.67	0.87 0.96	4.17 4.56				
HW-TSC	16.82	1.44 1.86	0.71 0.81	1.96 2.29	21.03	2.37 2.85	0.85 0.97	2.89 3.29	22.56	3.18 3.68	0.91 1.03	3.61 4.05
Xiaomi	23.75	2.04 3.61	0.82 1.28	2.62 4.78	24.34	2.97 6.48	0.90 2.11	3.71 9.86	24.56	3.87 8.55	0.95 2.28	4.29 11.15
Segmentation 1												
AISP-SJTU					19.18	2.84 3.16	0.87 0.94	4.94 5.38				
HW-TSC	14.44	1.53 1.98	0.68 0.76	2.42 2.76	17.63	2.64 3.14	0.82 0.91	3.50 3.92	18.85	3.66 4.18	0.89 0.99	4.37 4.84
Xiaomi	19.79	2.30 4.03	0.79 1.19	3.20 5.43	20.29	3.53 7.62	0.89 1.97	4.57 11.32	20.47	4.60 9.72	0.94 2.09	5.25 12.54
Segmentation 2												
AISP-SJTU					21.61	3.71 4.08	0.88 0.94	8.70 9.35				
HW-TSC	11.56	1.20 1.77	0.50 0.57	2.05 2.42	18.00	2.17 2.88	0.68 0.76	3.25 3.76	20.37	3.17 3.99	0.77 0.86	4.33 4.96
Xiaomi	22.15	1.85 4.50	0.69 1.19	3.04 8.10	22.71	3.23 8.80	0.77 2.10	4.84 18.63	23.08	4.43 11.55	0.83 2.30	5.63 21.16

[58] https://iwslt.org/2022/simultaneous

Human Evaluation Results

English-Japanese	BLEU	Error score	#Critical	#Major	#Minor
CUNI-KIT (high)	19.43	219	0	31	64
CUNI-KIT (low)	18.29	225	0	31	70
HW-TSC (medium)	15.21	472	2	85	37
NAIST (medium)	11.49	628	12	109	23

Table 16: Human evaluation results on one talk in the English-to-Japanese Simultaneous speech-to-speech translation task. Error weights are 5 for Critical and Major errors and 1 for Minor errors.

	Common	Non-native
Number of distinct audios	17	43
Mean length of audio (secs)	886	209
Total of subtitled audios annotated	439	1159
Mean ratings per annotated audio	164.4	40.8

Table 17: Human evaluation for the English-to-German task on two test sets: the Common one used also in automatic scoring and Non-native one. We show the size of the evaluation corpus, and the number of ratings collected.

		Common			Non-native		
	System	Low	Medium	High	Low	Medium	High
Human	CUNI-KIT	**3.13**	3.26	**3.44**	**2.46**	**2.57**	**2.98**
	UPV	2.96	**3.32**	3.40	2.07	2.55	2.72
	FBK	2.23	3.02	**3.44**	1.76	2.20	2.36
	HW-TSC	2.34	2.60	2.60	1.58	1.81	1.69
	NAIST	2.28	2.31	2.44	1.77	1.64	1.60
	Average±Std.dev.	2.59±0.38	2.90±0.39	3.06±0.45	1.93±0.31	2.15±0.38	2.27±0.55
	Interpreting		2.99			3.22	
BLEU	CUNI-KIT	**20.56**	**23.31**	**24.11**	**16.64**	**22.89**	**25.65**
	UPV	16.09	19.94	23.55	13.59	21.16	22.90
	FBK	10.23	20.12	23.59	8.40	16.51	20.42
	HW-TSC	13.97	19.10	19.73	10.35	13.47	13.55
	NAIST	13.40	15.29	15.47	6.11	9.33	9.25

Table 18: Human evaluation results for English-to-German Simultaneous task (upper part), compared with automatic BLEU scores (lower part). We calculate a mean score for each annotated audio file, then take the mean across all annotated audio files, for each system-latency combination. We highlight the best results in bold and report also the average across all submissions of a given latency band. The final row shows the results for human simultaneous interpreting (transcribed). The lower part reports the BLEU scores for the gold segmentation of the Common part of the test set (reported already on page 44) and for the Non-native part of the test set.
The BLEU scores correlate very well with the human judgement for each of the test sets parts: Pearson correlation across the systems and latency regimes is .898 for the Common part and .933 for the Non-native part. When considered together, the correlation decreases to .858.

B.2. Offline Speech Translation

Automatic Evaluation Results

Speech Translation: TED English-German tst 2022

· Systems are ordered according to `BLEU_NewRef`: *BLEU* score computed on the NEW reference set (literal translations).
· *BLEU* scores are given as percent figures (%).

System	BLEU_NewRef	BLEU_TEDRef	BLEU_MultiRef
USTC-NELSLIP cascade	26.7	23.9	37.6
YI end2end	25.7	23.6	36.5
YI cascade	25.6	23.7	36.4
USTC-NELSLIP end2end	25.3	22.9	35.7
NEMO	24.7	22.3	34.8
HW-TSC	24.2	20.8	33.5
KIT	23.9	22.0	33.8
FBK	23.6	21.0	32.9
UPC	23.0	20.8	32.3
ALEXA AI	22.6	20.1	31.5

Table 19: Official results of the **automatic evaluation** for the Offline Speech Translation Task, English to German.

Speech Translation: TED English-German tst 2021

· Systems are ordered according to `BLEU_TEDRef`: *BLEU* score computed on the ORIGINAL reference set.
· *BLEU* scores are given as percent figures (%).
· End-to-end systems are indicated by gray background.

System	BLEU_NewRef	BLEU_TEDRef	BLEU_MultiRef
USTC-NELSLIP cascade	28.9	24.1	40.3
YI cascade	28.1	23.2	39.0
YI end2end	27.8	23.1	38.8
HW-TSC	27.5	21.2	36.9
USTC-NELSLIP end2end	27.2	23.0	38.4
FBK	25.5	21.3	35.6
KIT	24.7	22.4	36.2
last Year's best	24.6	20.3	34.0
UPC	24.5	20.9	34.8
ALEXA AI	24.4	20.6	34.5

Table 20: Progress test set results of the **automatic evaluation** for the Offline Speech Translation Task, English to Japanese.

Speech Translation: TED English-Chinese tst 2022

· Systems are ordered according to BLEU_TEDRef: *BLEU* score computed on the ORIGINAL reference set.
· *BLEU* scores are given as percent figures (%).
· End-to-end systems are indicated by gray background.

System	BLEU_NewRef	BLEU_TEDRef	BLEU_MultiRef
USTC-NELSLIP cascade	35.8	35.7	44.1
YI cascade	34.7	35.0	42.9
HW-TSC	34.6	33.4	42.1
YI end2end	34.1	34.6	42.3
USTC-NELSLIP end2end	33.8	34.1	41.9
NEMO	33.3	33.7	41.2
NIUTRANS	32.3	33.2	40.5
KIT	31.1	32.0	39.0
ALEXA AI	30.4	30.8	37.9
UPC	29.2	29.9	36.4
NEURAL.AI	22.8	23.0	28.2

Table 21: Official results of the **automatic evaluation** for the Offline Speech Translation Task, English to Chinese.

Speech Translation: TED English-Japanese tst 2022

· Systems are ordered according to BLEU_TEDRef: *BLEU* score computed on the ORIGINAL reference set.
· *BLEU* scores are given as percent figures (%).
· End-to-end systems are indicated by gray background.

System	BLEU_NewRef	BLEU_TEDRef	BLEU_MultiRef
HW-TSC	22.7	14.3	30.8
USTC-NELSLIP cascade	21.6	20.1	33.4
USTC-NELSLIP end2end	20.5	17.4	30.5
YI end2end	18.0	19.1	29.8
YI cascade	18.7	20.2	31.3
KIT	16.2	17.2	26.4
UPC	15.1	15.6	24.7
ALEXA AI	15.3	16.2	25.3

Table 22: Official results of the **automatic evaluation** for the Offline Speech Translation Task, English to Japanese.

Human Evaluation Results

Speech Translation: TED English-German tst 2022 (subset)

Rank	Ave.	Ave. z	System
1-3	88.9	0.142	USTC-NELSLIP cascade
1-4	87.4	0.075	USTC-NELSLIP end2end
1-4	87.6	0.063	YI cascade
4-9	86.5	0.008	KIT
4-9	86.1	-0.004	FBK
2-7	86.3	-0.011	YI end2end
4-9	85.6	-0.023	NEMO
5-9	85.4	-0.039	UPC
5-9	84.8	-0.076	HW-TSC
10	83.9	-0.133	ALEXA AI

Table 23: Official results of the **human evaluation** for the Offline Speech Translation Task, English to German. Systems ordered by the standardized DA z-score. Systems within clusters indicated by horizontal lines are considered tied. Scores collected using direct assessment with previous/next-sentence context.

Speech Translation: TED English-Chinese tst 2022 (subset)

1	85.6	0.184	USTC-NELSLIP cascade
2-5	84.2	0.121	YI end2end
2-7	84.0	0.097	YI cascade
2-7	83.5	0.086	USTC-NELSLIP end2end
3-8	83.1	0.061	NEMO
3-8	83.2	0.057	KIT
2-7	82.8	0.038	HW-TSC
6-9	82.4	0.023	NIUTRANS
8-10	81.6	-0.023	ALEXA AI
9-10	80.8	-0.055	UPC
11	71.2	-0.589	NEURAL.AI

Table 24: Official results of the **human evaluation** for the Offline Speech Translation Task, English to Chinese. Systems ordered by the standardized DA z-score. Systems within clusters indicated by horizontal lines are considered tied. Scores collected using direct assessment with previous/next-sentence context.

Speech Translation: TED English-Japanese tst 2022 (subset)

1-4	78.4	0.086	YI cascade
1-4	77.6	0.065	USTC-NELSLIP cascade
1-4	77.6	0.061	YI end2end
1-4	76.6	0.005	HW-TSC
5-6	76.3	-0.009	USTC-NELSLIP end2end
5-6	76.3	-0.013	KIT
7-8	74.7	-0.082	ALEXA AI
7-8	73.2	-0.113	UPC

Table 25: Official results of the **human evaluation** for the Offline Speech Translation Task, English to Japanese. Systems ordered by the standardized DA z-score. Systems within clusters indicated by horizontal lines are considered tied. Scores collected using direct assessment with previous/next-sentence context.

B.3. Speech to Speech Translation

Results for the speech to speech translation task, described in Section 4.

While both automatic metrics and human evaluation are provided, the task ranking was determined by human evaluation of translation quality (Table 28).

System	BLEU	chrF
MLLP-VRAIN	**19.70**	53.15
HW-TSC primary	19.58	**53.81**
HW-TSC contrastive3	19.35	53.75
HW-TSC contrastive1	19.22	53.65
HW-TSC contrastive2	18.90	53.00
UPC	16.38	50.20
Reference text (+TTS)	68.46	88.78
FBK Offline (+TTS)	17.37	51.21
KIT Offline (+TTS)	16.63	50.43
Reference text (+normalization)	100.00	100.00
FBK Offline (+normalization)	23.44	55.84
KIT Offline (+normalization)	23.51	55.18

Table 26: **S2ST: automatic metrics.** Speech output is first transcribed with ASR before scoring against reference text. Text is normalized for scoring (punctuation and case removed, whitespace standardized). The effects of synthesis + ASR transcription are shown by synthesizing the reference text and selected Offline task submissions and scoring after ASR.

System	nat.	clar.	sound.
MLLP-VRAIN	4.156 (0.037)	4.626 (0.028)	4.562 (0.028)
HW-TSC primary	3.135 (0.042)	3.835 (0.037)	3.867 (0.034)
UPC	3.118 (0.042)	3.786 (0.037)	3.862 (0.032)
Reference	3.116 (0.043)	3.678 (0.038)	3.799 (0.032)

Table 27: **S2ST: speech quality human evaluation.** System outputs were evaluated along 3 dimensions, which are more fine-grained than mean opinion score: speech naturalness (nat.), clarity of speech (clar.) and sound quality (sound.). Numbers in parenthesis indicate a 95% confidence interval.

System	Translation quality
HW-TSC primary	4.606 (0.034)
MLLP-VRAIN	4.439 (0.057)
UPC	4.374 (0.041)
Reference	4.369 (0.038)

Table 28: **S2ST: translation quality human evaluation.** The initial MLLP-VRAIN submission had a misalignment and was later fixed. As a result, the number of samples for MLLP-VRAIN is 1000 instead of 2059. Numbers in parenthesis indicate a 95% confidence interval.

B.4. Dialect Speech Translation

Automatic Evaluation Results

Tunisian Arabic→English

Team	Condition	System	test2					test1
			BLEU↑	BP	pr1	chrF2	TER↓	BLEU
CMU	dialect adapt	primary (E2)	20.8 ± 0.7	0.931	53.1	44.3	64.5	19.5
CMU	dialect adapt	contrastive	20.7 ± 0.7	0.929	53	44.1	64.6	19.3
CMU	basic	primary (E1)	20.4 ± 0.7	0.944	52.2	43.8	65.4	19.2
CMU	basic	contrastive	20.1 ± 0.7	0.936	52.2	43.5	65.3	19
CMU	dialect adapt	contrastive (D6)	19.8 ± 0.7	0.902	53.2	43.3	64.6	18.9
CMU	basic	contrastive (D3)	19.7 ± 0.7	0.916	52.4	43	65.5	18.7
CMU	dialect adapt	contrastive (D5)	19.5 ± 0.6	0.896	53.2	42.8	64.6	18.3
CMU	dialect adapt	contrastive (C6)	19.4 ± 0.6	0.937	50.7	43	67.1	17.9
CMU	basic	contrastive (D2)	19.1 ± 0.6	0.939	51.3	42.7	66.5	18.1
JHU	dialect adapt	primary	18.9 ± 0.7	0.99	48	42.1	70.2	17.8
JHU	unconstrain.	primary	18.7 ± 0.7	0.959	48.7	41.6	69.2	17.5
CMU	basic	contrastive (C3)	18.6 ± 0.6	0.942	49.4	41.8	68.3	17.5
JHU	basic	primary	17.1 ± 0.6	0.973	46.8	40.4	71.4	16.1
ON-TRAC	unconstrain.	post-evaluation	14.4 ± 0.6	1	42.7	36.5	76.7	-
ON-TRAC	unconstrain.	contrastive1	13.6 ± 0.6	1	41.7	35.7	78.3	-
ON-TRAC	basic	primary	12.4 ± 0.6	0.8	44.3	32.8	75.5	-
ON-TRAC	unconstrain.	contrastive2	11.3 ± 0.5	0.95	38.7	32.7	80.6	-
Baseline	basic	baseline E2E	11.1 ± 0.5	0.885	40	31.9	77.8	10.1

Table 29: Automatic evaluation results for the Dialect Speech Translation Task. Systems are ranked in order of the official metric: BLEU on test2 blind evaluation set. We also report chrF2, TER, as well as the brevity penalty (BP) and 1-gram precision (pr1) components of BLEU. We further use bootstrap resampling (1k samples) and report the 95% confidence interval for BLEU on test2 (Koehn, 2004). For details of each system, refer to the system name in the respective papers.

Tunisian Arabic ASR Automatic Evaluation Results

ASR System	WER↓		CER↓	
	Orig	Norm	Orig	Norm
JHU / basic / primary	70.5	43.8	30.5	22.5
JHU / dialect adapt / primary	70.1	42.9	30.4	22.3
JHU / unconstrained / primary	69.4	42.8	30.6	22.5
ON-TRAC / unconstrained / primary	68.2	45.1	28.4	21.5
ON-TRAC / unconstrained / post-eval	65.7	41.5	28.1	21.1

Table 30: Word Error Rate (WER) and Character Error Rate (CER) of the ASR component of submitted cascaded systems on test2. This is computed by comparing ASR hypotheses with the Tunisian manual transcripts. The original version (Orig) matches the minimal text pre-processing provided by the organizer's data preparation scripts, and results in relatively high WER. Transcription standards for primarily spoken dialects are challenging, so it may be beneficial as diagnosis to run some additional Arabic-specific normalization (Norm) for e.g. Alif, Ya, Ta-Marbuta on the hypotheses and transcripts before computing WER/CER. We are grateful to Ahmed Ali for assistance on this.

Human Evaluation Results

Tunisian Arabic→English

Rank	Ave.	Ave. z	Team / Condition / System
1	76.6	0.457	translator-A
2-3	66.5	0.119	CMU / dialect adapt / contrastive (D6)
2-3	66.5	0.114	CMU / dialect adapt / primary (E2)
4-5	62.7	-0.032	JHU / dialect adapt / primary
4-5	60.7	-0.093	JHU / basic condition / primary
6-7	56.1	-0.271	ON-TRAC / unconstrained / primary
6-7	55.3	-0.302	ON-TRAC / unconstrained / contrastive1

Table 31: Official results of the human evaluation for the Dialect Speech Translation Task. Systems ordered by the standardized DA z-score. Systems within clusters indicated by horizontal lines are considered tied. Scores collected using the document-level DA+SQM task in Appraise.

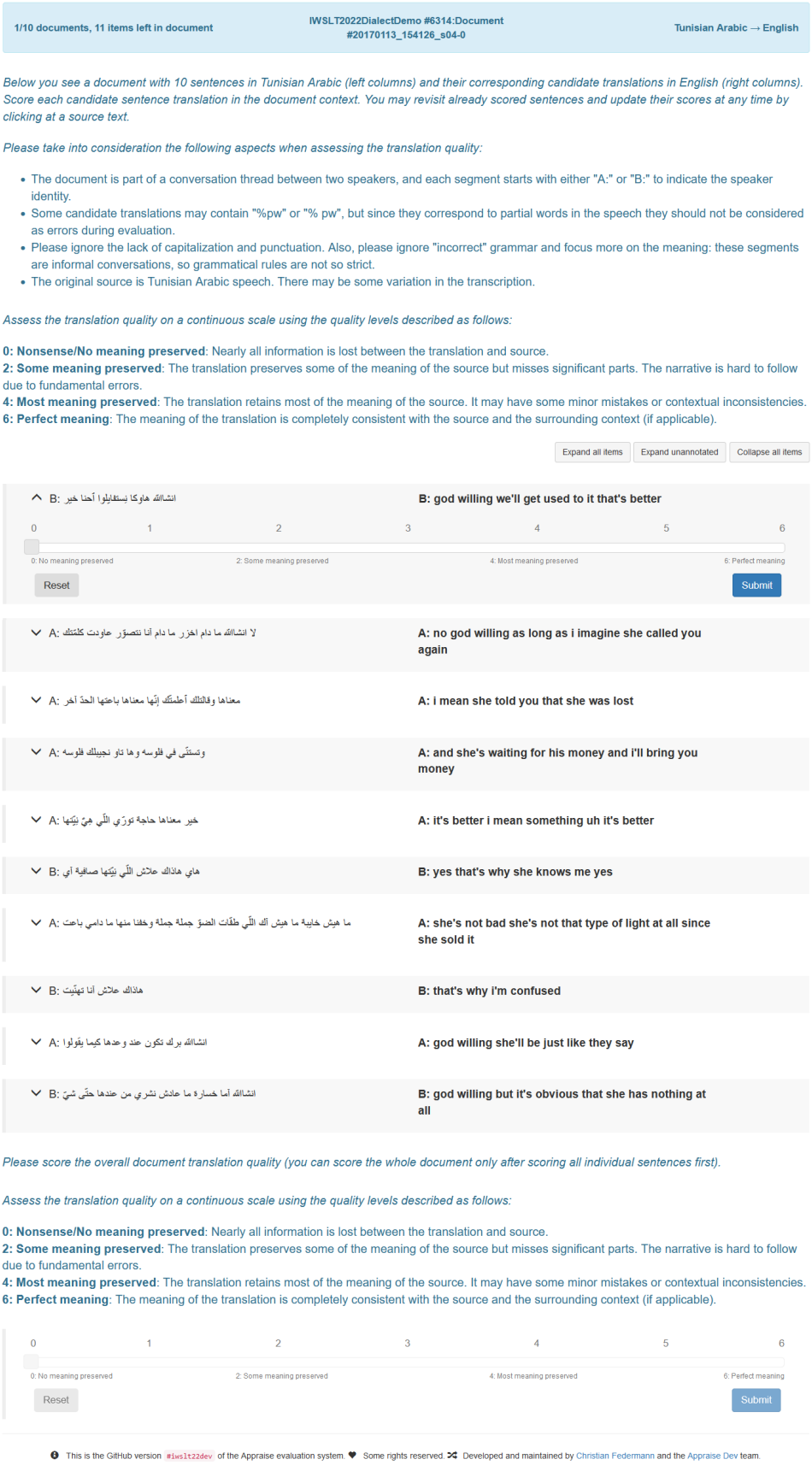

Figure 6: A screen shot of an example annotation task in Appraise featuring source-based document-level Direct Assessment with SQM for the Dialect Speech Translation Task.

B.5. Formality Control For Speech Translation

Automatic Evaluation Results

		EN→HI		EN→JA		EN→DE		EN→ES		EN→IT		EN→RU	
Setting	System	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
unconstrained	baseline	22.0	0.67	17.9	0.24	32.6	0.55	37.4	0.70	32.2	0.64	19.5	0.32
	ALEXA AI	38.9	0.874	19.4	0.378								
	UMD	12.1	0.192	11.6	-0.023	22.4	0.161	27.8	0.344	22.9	0.247	14.4	0.075
	UOS					32.5	0.497	37.0	0.635	33.1	0.562	21.5	0.357
constrained	UOS					31.5	0.448	36.5	0.608	33.1	0.553	21.4	0.329

Table 32: Automatic evaluation using sacrebleu and COMET on generic test sets. For EN→DE, ES, IT, RU participants were asked to evaluated their systems on MuST-C dataset. We have also included baseline models trained in the *unconstrained* setting for comparison. For EN→HI, JA participants were evaluated on WMT Newstest 2014 and 2020 respectively.

		Supervised								Zero-shot			
		EN→HI		EN→JA		EN→DE		EN→ES		EN→IT		EN→RU	
Setting	System	F	I	F	I	F	I	F	I	F	I	F	I
unconstrained	baseline (generic)	96.3	3.70	49.6	50.3	45.8	54.2	36.6	63.4	3.70	94.5	93.4	6.60
	ALEXA AI	99.6	99.8	88.8	98.8								
	UMD	99.4	98.7	86.3	97.5	99.4	96.5	99.5	93.2	32.8	97.9	100.0	1.10
	UOS					100.0	100.0	98.1	100.0	51.2	98.6	99.5	85.8
constrained	UOS					100.0	88.6	87.4	98.0	29.5	92.9	98.1	15.4

Table 33: Automatic evaluation of formality control accuracy (M-ACC) reported for Formal (F) and Informal (I). For comparison, we have included our baseline generic (uncontrolled) performance on the formality testset. For EN→IT, RU participants were given a zero-shot task and asked to train a formality controlled model without labelled training data in Italian or Russian.

Human Evaluation Results

Lang.	Setting	Sys.	Control	F	I	N	O	IAA
EN→JA	unconstrained	UMD	Formal	89.3	0.7	0.0	9.7	0.90
		UMD	Informal	2.0	92.5	0.0	5.5	
		ALEXA AI	Formal	82.8	1.3	0.0	15.5	
		ALEXA AI	Informal	3.0	82.7	0.0	14.3	
EN→IT	unconstrained	UMD	Formal	13.7	25.2	47.0	14.2	0.91
		UMD	Informal	1.0	78.3	11.5	9.2	
		UOS	Formal	6.0	7.2	81.3	5.5	
		UOS	Informal	0.3	81.0	13.2	5.5	
	constrained	UOS	Formal	0.2	10.2	87.7	2.0	
		UOS	Informal	0.2	36.3	58.3	5.2	
EN→RU	unconstrained	UMD	Formal	77.2	0.2	7.0	15.7	0.85
		UMD	Informal	74.3	0.7	7.8	17.2	
		UOS	Formal	85.0	0.3	6.0	8.7	
		UOS	Informal	10.3	71.3	3.2	15.2	
	constrained	UOS	Formal	85.3	2.0	5.7	7.0	
		UOS	Informal	65.0	12.7	6.3	16.0	

Table 34: Percentage of system outputs (with a given formality level (Control) and setting (Setting)) labeled by professional translators according to the formality level: formal (F), informal (I), neutral (N), other (O). IAA was computed using the Krippendorff's α coefficient.

B.6. Isometric Spoken Language Translation

Automatic MT Evaluation Results

	En→De		
System	BERTScore	LC	BLEU(detok)
StrongBaseline*	**77.44**	68.0	**21.6**
AppTek-Constrained	77.32	86.5	18.7
HW-TSC-Unconstrained	75.79	96.5	20.2
APV-Unconstrained	73.68	39.0	16.5
WeakBaseline	74.86	43.0	15.5
HW-TSC-Constrained	74.07	**98.0**	17.9

	En→Fr		
System	BERTScore	LC	BLEU(detok)
StrongBaseline*	**81.75**	75.5	**36.2**
NUV-Unconstrained	79.96	47.5	27.1
APV-Unconstrained	77.77	45.0	32.9
HW-TSC-Constrained	76.11	**96.0**	31.5
WeakBaseline	77.18	37.0	25.2

	En→Es		
System	BERTScore	LC	BLEU(detok)
StrongBaseline*	**81.86**	80.5	**36**
APV-Unconstrained	80.87	49.5	35.3
HW-TSC-Constrained	78.57	**96.5**	29.9
WeakBaseline	78.32	51.0	27.7

Table 35: Automatic evaluation results for Isometric SLT task on the blind test set. Metrics are computed using the submissions primary system. System ranking follows the human evaluation ranking in Table 36. If BERTScore is a tie, system with the highest LC wins (*). BERTSCore and LC are the primary metrics for the task, detoknized-BLEU is provided only as a secondary reference. ***Bold*** *highlights the top score*.

MT Human Evaluation Results

En→De

Rank	Ave.	Ave. z	System
1	89.0	0.755	translator-A
2-3	72.6	0.189	STRONGBASELINE
2-3	69.9	0.123	APPTEK-Constrained
4-5	62.6	-0.153	HW-TSC-Unconstrained
4-6	62.1	-0.224	APV-Unconstrained
5-7	59.4	-0.298	WEAKBASELINE
6-7	56.3	-0.467	HW-TSC-Constrained

En→Fr

Rank	Ave.	Ave. z	System
1	80.8	0.624	translator-A
2-3	64.3	0.009	STRONGBASELINE
2-4	60.2	-0.152	NUV-constrained
3-6	58.0	-0.280	APV-Unconstrained
4-6	53.2	-0.348	HW-TSC-Constrained
4-6	53.6	-0.389	WEAKBASELINE

En→Es

Rank	Ave.	Ave. z	System
1	82.5	0.601	translator-A
2-3	70.3	0.020	STRONGBASELINE
2-3	69.9	-0.031	APV-Unconstrained
4-5	64.0	-0.283	HW-TSC-Constrained
4-5	59.8	-0.409	WEAKBASELINE

Table 36: Official results of the text-based human evaluation for the Isometric SLT Task. Systems ordered by the standardized DA z-score. Systems within clusters indicated by horizontal lines are considered tied. Scores collected using the document-level DA+SQM task in Appraise.

Automatic Dubbing Human Evaluation Results

En→De

Comparison	Wins (%)
WEAKBASELINE vs APPTEK-Constrained	32.9 vs 49.8*
WEAKBASELINE vs HW-TSC-Constrained	29.0 vs 49.4*
WEAKBASELINE vs HW-TSC-Unconstrained	41.1 vs 44.2
WEAKBASELINE vs APV-Unconstrained	37.9 vs 42.5
WEAKBASELINE vs STRONGBASELINE	29.0 vs 52.3*
APPTEK-Constrained vs HW-TSC-Constrained	42.4 vs 38.8
APPTEK-Constrained vs HW-TSC-Unconstrained	41.0 vs 38.0
APPTEK-Constrained vs APV-Unconstrained	43.9 vs 36.9
APPTEK-Constrained vs STRONGBASELINE	38.0 vs 39.6
HW-TSC-Constrained vs HW-TSC-Unconstrained	38.3 vs 36.0
HW-TSC-Constrained vs APV-Unconstrained	44.3 vs 37.7
HW-TSC-Constrained vs STRONGBASELINE	36.0 vs 42.7
HW-TSC-Unconstrained vs APV-Unconstrained	49.3 vs 32.7*
HW-TSC-Unconstrained vs STRONGBASELINE	37.2 vs 41.8
APV-Unconstrained vs STRONGBASELINE	31.3 vs 49.7*

En→Fr

Comparison	Wins (%)
WEAKBASELINE vs HW-TSC-Constrained	31.7 vs 51.7*
WEAKBASELINE vs NUV-Unconstrained	32.6 vs 50.9*
WEAKBASELINE vs APV-Unconstrained	25.7 vs 55.7*
WEAKBASELINE vs STRONGBASELINE	26.7 vs 57.0*
HW-TSC-Constrained vs NUV-Unconstrained	40.0 vs 40.0
HW-TSC-Constrained vs APV-Unconstrained	46.7 vs 34.7+
HW-TSC-Constrained vs STRONGBASELINE	31.9 vs 49.1*
NUV-Unconstrained vs APV-Unconstrained	35.6 vs 40.0
NUV-Unconstrained vs STRONGBASELINE	29.0 vs 48.6*
APV-Unconstrained vs STRONGBASELINE	34.3 vs 44.7

En→Es

Comparison	Wins (%)
WEAKBASELINE vs HW-TSC-Constrained	21.0 vs 51.0*
WEAKBASELINE vs APV-Unconstrained	30.3 vs 46.7*
WEAKBASELINE vs STRONGBASELINE	24.3 vs 53.7*
HW-TSC-Constrained vs APV-Unconstrained	37.7 vs 35.7
HW-TSC-Constrained vs STRONGBASELINE	34.3 vs 40.0
APV-Unconstrained vs STRONGBASELINE	30.3 vs 44.7*

Table 37: Automatic dubbing human evaluation results on pairwise comparisons of submitted systems for the Isometric SLT task. We report the Wins, i.e, the % of times one condition is preferred over the other with statistical significance levels $p < 0.01$(*) and $p < 0.05$(+).

En→De

Rank	N_{Wins}	System
1	5	STRONGBASELINE
2	4	APPTEK-Constrained
3	3	HW-TSC-Constrained
4	2	HW-TSC-Unconstrained
5	1	APV-Unconstrained
6	0	WEAKBASELINE

En→Fr

Rank	N_{Wins}	System
1	4	STRONGBASELINE
2	2	HW-TSC-Constrained
3	2	APV-Unconstrained
4	1	NUV-Constrained
5	0	WEAKBASELINE

En→Es

Rank	N_{Wins}	System
1	3	STRONGBASELINE
2	2	HW-TSC-Constrained
3	1	APV-Unconstrained
4	0	WEAKBASELINE

Table 38: Results of human evaluation of dubbed videos. Systems are ranked using N_{Wins}, i.e., the number of evaluations for which that systems was preferred over some other system.

En→De

Systems	Smoothness	LC
STRONGBASELINE	88.55	68
APPTEK-Constrained	86.22	86.5
HW-TSC-Constrained	88.45	98
HW-TSC-Unconstrained	88.92	96.5
APV-Unconstrained	82.53	39
WEAKBASELINE	84.22	43

En→Fr

Systems	Smoothness	LC
STRONGBASELINE	80.66	75.5
HW-TSC-Constrained	77.93	96
APV-Unconstrained	78.31	45
NUV-Constrained	75.52	47.5
WEAKBASELINE	66.84	37

En→Es

Systems	Smoothness	LC
STRONGBASELINE	92.01	80.5
HW-TSC-Constrained	92.65	96.5
APV-Unconstrained	92.02	49.5
WEAKBASELINE	85.21	51

Table 39: Results of automatic evaluation for subset of 60 dialogues used for dubbing evaluation using smoothness (Federico et al., 2020a) that measures the stability of speaking rate across contiguous phrases and length compliance (LC).

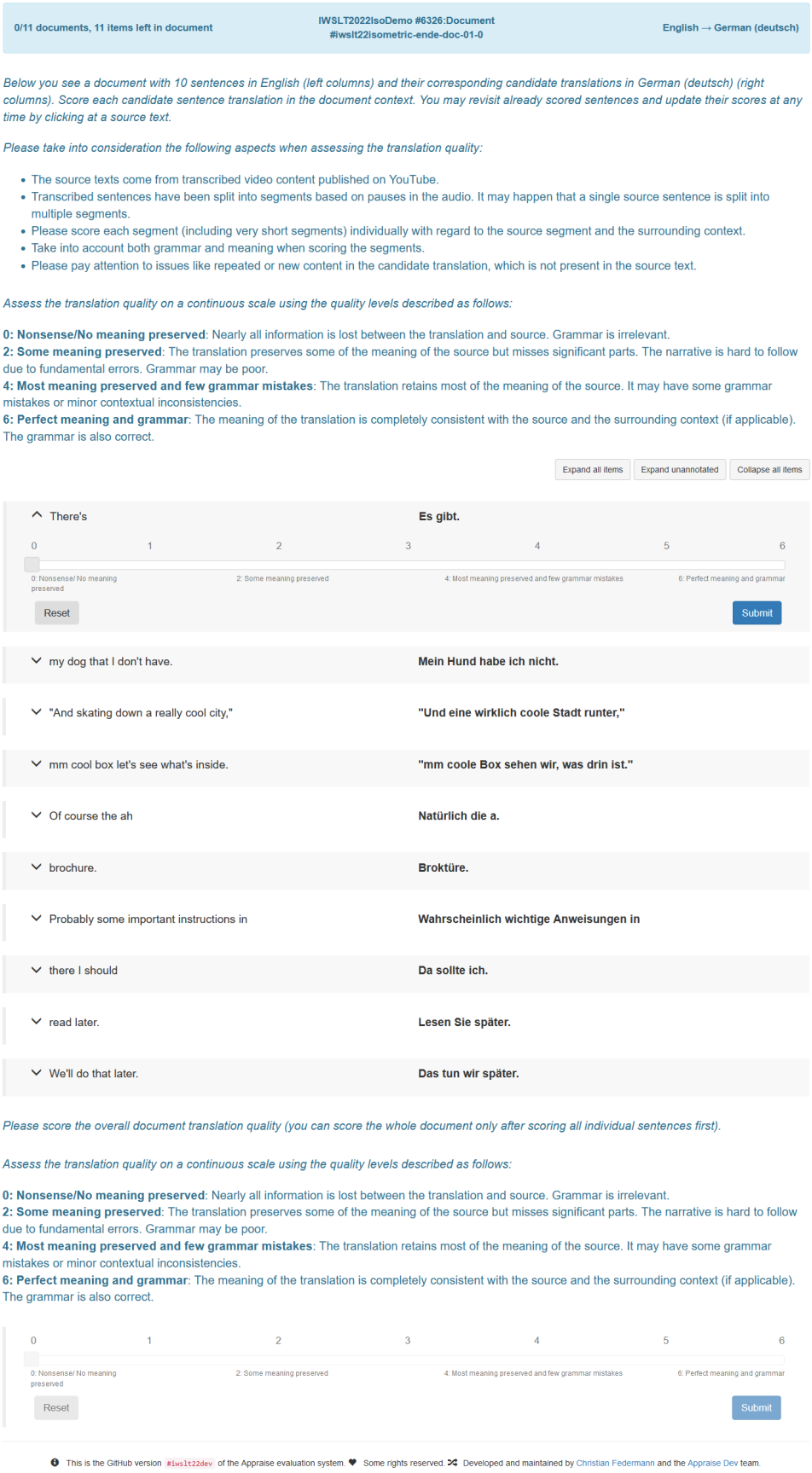

Figure 7: A screen shot of an example annotation task in Appraise featuring source-based document-level Direct Assessment with SQM for the Isometric SLT Task.