Evaluating the IWSLT2023 Speech Translation Tasks: Human Annotations, Automatic Metrics, and Segmentation

Matthias Sperber[1], **Ondřej Bojar**[2], **Barry Haddow**[3], **Dávid Javorský**,[2]
Xutai Ma[4], **Matteo Negri**[5], **Jan Niehues**[6], **Peter Polák**[2]
Elizabeth Salesky[1,7], **Katsuhito Sudoh**[8], **Marco Turchi**[9]
[1]Apple, [2]Charles U., [3]U. Edinburgh, [4]Meta, [5]FBK, [6]KIT, [7]JHU, [8]NAIST, [9]Zoom
sperber@apple.com, bojar@ufal.mff.cuni.cz, bhaddow@ed.ac.uk, javorsky@ufal.mff.cuni.cz,
xutaima@meta.com, negri@fbk.eu, jan.niehues@kit.edu, polak@ufal.mff.cuni.cz
esalesky@jhu.edu, sudoh@is.naist.jp, marco.turchi@zoom.us

Abstract

Human evaluation is a critical component in machine translation system development and has received much attention in text translation research. However, little prior work exists on the topic of human evaluation for speech translation, which adds additional challenges such as noisy data and segmentation mismatches. We take first steps to fill this gap by conducting a comprehensive human evaluation of the results of several shared tasks from the last International Workshop on Spoken Language Translation (IWSLT 2023). We propose an effective evaluation strategy based on automatic resegmentation and direct assessment with segment context. Our analysis revealed that: *1)* the proposed evaluation strategy is robust and scores well-correlated with other types of human judgements; *2)* automatic metrics are usually, but not always, well-correlated with direct assessment scores; and *3)* COMET as a slightly stronger automatic metric than chrF, despite the segmentation noise introduced by the resegmentation step systems. We release the collected human-annotated data in order to encourage further investigation.

Keywords: Human evaluation, speech translation, evaluation metrics

1. Introduction

Human evaluation plays a critical role in the development of translation systems and, although costly, it serves as a gold standard against which to calibrate automatic metrics.[1] Moreover, it is important when in doubt about whether certain types of system behaviors are being accurately captured by automatic metrics. This is particularly relevant when the output quality is high, and automatic metrics become less reliable. For the case of machine translation for *text* inputs (MT), a large body of prior research has investigated and improved procedures for manual evaluation, for example in the Conference on Machine Translation (WMT) series of shared tasks (Kocmi et al., 2022). Many of the popular MT metrics have been shown to have a high correlation with human evaluation results (Freitag et al., 2022).

In this paper, we focus on evaluation in *speech* translation (ST), which introduces additional complicating factors, such as erroneous automatic segmentation, dealing with noisy inputs, conversational and disfluent language, and special downstream requirements, such as simultaneous translation, and subtitling/dubbing constraints. To our knowledge, little prior work has focused on meta-analysis of human evaluation for speech translation. Consequently, it remains unclear to what degree procedures for human and automatic evaluation from MT can be transferred to the ST scenario.

As a first step toward developing trusted human evaluation procedures for ST, and following the successful related efforts at WMT, we conducted a manual evaluation of several shared tasks from the International Workshop for Spoken Language Translation (IWSLT) 2023. Tasks include translation of presentations at the Conference of the Association of Computational Linguistics (ACL) and TED talks[2] in several language pairs in offline, multilingual, and simultaneous translation conditions. In the offline and multilingual conditions, inputs are given as unsegmented long-form speech, requiring systems to apply automatic segmentation prior to translation. The shared tasks used test sets that partially overlap between different task conditions, providing the opportunity for analysis across various conditions.

For conducting our human evaluation, we choose an approach that allows the comparison of systems despite potentially mismatching segmentation by re-segmenting system outputs to a common segmentation and using segment context. To minimize costs, a random subset of segments is chosen for evaluation.

Through extensive analysis, we shed light on

[1]This is especially true for *trainable* metrics, which have recently started to outperform traditional metrics (Freitag et al., 2022).

[2]Talks on technology, entertainment, and design, available at www.ted.com.

LREC-COLING 2024, pages 6484–6495

Task	Offline	Multilingual	Simultaneous
TED	✓		✓
ACL	✓	✓	

Table 1: Domains used by the 3 shared tasks.

the current state of ST evaluation protocols contributing in three ways. First, we confirm that our collected direct assessment (DA) scores are well-correlated with additionally collected annotations, namely Multidimensional Quality Metric (MQM) and continuous ratings, indicating that the proposed evaluation procedure is sound and reliable to be used in future work. Second, we show that the correlation between human evaluation and automatic metrics is often, but not always, high. We conclude that these metrics can generally be trusted, but that human evaluation should be continually run side-by-side for further verification. Third, we show that COMET (Rei et al., 2020) has higher correlation than other automatic metrics despite potential robustness issues due to noisy automatic segmentation. This shows promise in using trainable metrics like COMET for speech translation scenarios, although more detailed follow-up experiments is needed. To the best of our knowledge, the released annotations would be the first publicly available annotations of their kind for speech rather than text translation, facilitating the comparison of evaluation methodology between modalities.[3]

2. Task Description

In this section, we provide a bird's eye view of the IWSLT 2023 tasks considered in our study. Accordingly, we concentrate on the *task conditions*, the *data* used, and the *automatic evaluation* protocols.

2.1. Conditions

2.1.1. Offline

The Offline Speech Translation Task aimed to explore automatic methods for translating spoken language in one language into written text in another language. This could be achieved through either cascaded solutions involving pipelined automatic speech recognition and machine translation (MT) systems as core components, or end-to-end approaches that directly translate the audio bypassing the intermediate transcription step. Recent results (Bentivogli et al., 2021) have shown that the performance of end-to-end models is becoming comparable to that of cascade solutions, but the best-performing technology has yet to be identified.

In the 2023 edition, the Offline Speech Translation Task not only addressed the question of whether the cascade solution remains the prevailing technology but it also assessed the systems submitted in more intricate and demanding situations. These included multiple speakers, non-native speakers, different accents, varying recording quality, specialized terminology, controlled interaction with a second speaker, and spontaneous speech. Submitted systems used a variety of approaches, experimenting with both constrained and unconstrained training data conditions as well as use of pretrained large language models.

2.1.2. Multilingual

The Multilingual Task focused particularly on the capability to translate to a wide variety of target languages in a realistic use case: with long-form audio requiring segmentation, diverse speaker accents, and domain-specific terminology. The task used the full set of ten target languages from the ACL 60-60 evaluation sets,[4] of which three are considered in this paper (see Section 2.2). Teams were required to submit to all language pairs, though not restricted to multilingual modelling approaches, such that all approaches could be compared across all ten target languages. There were comparisons between both cascaded and end-to-end systems, a variety of pretrained models, as well as multilingual and single language pair finetuning.

2.1.3. Simultaneous

Simultaneous translation, also known as real-time or online translation, is the task of generating translations incrementally given partial input only. It enables low-latency applications such as simultaneous interpretation in personal traveling and international conferences. In the 2023 edition, a submission qualified as a simultaneous system if the latency is no greater than average lagging (Ma et al., 2019) of 2 seconds.

The organizers of the shared task proposed two tracks, speech-to-text and speech-to-speech, over three language pairs, English to German, Chinese and Japanese. The human evaluation is only conducted over speech-to-text systems.

2.2. Test Data

As shown in Table 1, test data came from the TED domain (for offline and simultaneous tasks), and ACL domain (for offline and multilingual tasks), as detailed below. For the offline and multilingual task,

[3]Data is available in WMT format under https://huggingface.co/datasets/IWSLT/da2023.

[4]Arabic, Mandarin Chinese, Dutch, French, German, Japanese, Farsi, Portuguese, Russian, and Turkish.

this data was given as long-form speech, requiring systems to use automatic segmentation strategies. The simultaneous task provided input data according to the reference segmentation. For both datasets, our analysis concentrates on German, Japanese, and Mandarin Chinese as target language, and English as source language.

2.2.1. TED

For the TED scenario, the test sets are built starting from 42 talks that are not yet part of the en-de section of the current public release of the TED-derived MuST-C corpus (Cattoni et al., 2021). From this material, talks that have been translated into Japanese and Chinese were selected to build the en-ja and en-zh test sets, which consist of 37 and 38 talks respectively.

There are two different types of target-language references used in the evaluation campaign. The first type is the original TED translations, which come in the form of subtitles. Because of TED's subtitling guidelines,[5] these translations may contain compressed or omitted content, making them less literal compared to unconstrained translations. The second type is unconstrained translations, which were created from scratch by professionals, following typical translation guidelines. These translations are therefore exact, meaning they are literal and have proper punctuation.

2.2.2. ACL

The second test data set is based on the ACL 60-60 evaluation sets released by Salesky et al. (2023), which are composed of technical presentations from ACL 2022 given in English by diverse speakers across different paper topics. The talks were manually sentence segmented, transcribed, and translated into ten target languages from the 60/60 initiative.[6] The evaluation sets contain 1 hour of one-to-many parallel speech, transcripts, and translations per language pair.

The ACL test sets introduce additional challenges beyond those in the traditional TED setting, including the presence of non-native speakers with diverse accents, varying recording quality, and the strong presence of technical terminology.

3. Evaluation

3.1. Resegmentation

To follow realistic use conditions, no reference segmentation was provided for the offline and multilingual tasks; rather, participants were required to process each talk as a whole. To this end, participants used different tools (either publicly available or proprietary) to automatically segment the audio for downstream processing. Consequently, the segmentation in each submitted system can differ significantly from the reference segmentation. To parallelize outputs and references for evaluation, we re-segmented translation hypotheses following Matusov et al. (2005) by exploiting WER alignment to the reference translation with the tool mwerSegmenter.[7] In this step, the hypotheses are monotonically re-segmented in order to minimize the global WER to the reference translation, with candidate boundaries determined by tokenization; in our case, word-level for German and character-level for Chinese and Japanese. An illustrative example from the ACL data is shown in Figure 1. All subsequent evaluation, both automatic and human, used the re-segmented hypotheses which are now parallel to the references.

3.2. Automatic Metrics

For automatic evaluation, the evaluation campaign focused on standard lexical and model-based machine translation evaluation metrics, namely BLEU (Papineni et al., 2002), chrF (Popović, 2015) and COMET[8] (Rei et al., 2020). For conciseness, here we focus on the last two, as BLEU has been shown to be less reliable than chrF (Freitag et al., 2022). As is common practice in speech translation evaluation, automatic metrics are computed using the re-segmented hypotheses as described in Section 3.1.

3.3. Human Evaluation

Conducting human evaluation is important for a number of reasons. Among others, prior work (Barrault et al., 2019) has observed that for highly accurate translation systems, human evaluators often rank systems differently than automatic metrics, indicating that automatic metrics alone may not be reliable enough in many situations. Moreover, recent speech translation research places much focus on comparing two different modeling paradigms, i.e. direct vs. cascaded ST. This is reminiscent of an earlier situation in MT research, where automatic metrics, when used to compare different MT paradigms (rule-based, statistical, neural), were found to be less reliable than for comparing similar systems built under the same paradigm (Bojar et al., 2016). Similarly, it could be conceivable that automatic metrics may suffer from systematic biases that would make comparison across ST modeling paradigms difficult. However, whether (and to what

[5] https://www.ted.com/participate/translate/guidelines

[6] www.2022.aclweb.org/dispecialinitiative

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

[8] Unbabel/wmt22-comet-da

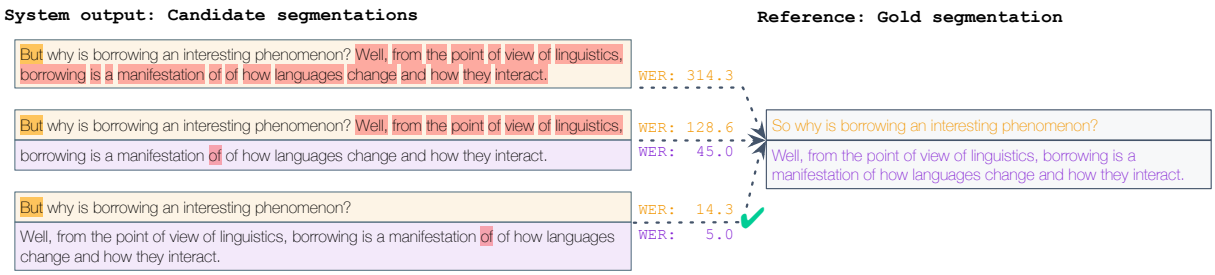

Figure 1: To parallelize system outputs with reference translations, we re-segment hypotheses by minimizing the overall WER to the reference with MWERSEGMENTER, which supports many-to-many realignment.

Source text:	**We assume the precision of quantities are known.**
Reference target text:	*We go under assumption, that the precision of the sets known is.* **Wir gehen davon aus dass die Genauigkeit der Mengen bekannt ist.**
Reseg. sys. output:	*we under assumption, that the precision of the sets known is,* **wir davon aus, dass die Präzision von Mengen bekannt ist,**
Previous sys. output:	*It holds thus also certain assumptions. As in prior works go* **Es gelten also auch bestimmte Annahmen. Wie in früheren Arbeiten gehen**
Next system output:	*and consider only basic operators like additon, subtraction, ...* **und betrachten nur grundlegende Operatoren wie Addition, Subtraktion, ...**

Figure 2: Example for annotation of re-segmented English-German system outputs. Gloss is provided in italics. The human annotator, if looking only at the re-segmented system output of the current sentence, would find the main verb missing, the beginning of the sentence not capitalized, and the sentences ending in a comma instead of full stop, and penalize the sentence accordingly. As a remedy, we also present system outputs for adjacent segments to the annotator. The main verb (*gehen*) is located in the system output for the previous sentence, and by being presented with the context, the annotator will be able to see that this translation has no grammatical issues nor problems with casing or punctuation.

degree) these modeling paradigms lead to systematically different translations is currently not well understood.[9] In light of this, we argue that more emphasis is needed on human evaluation for comparing ST paradigms. Beyond such considerations, setting up a gold procedure is important for a variety of reasons, including establishing or training better automatic metrics.

3.3.1. Direct assessment

We conduct a source-based DA (Graham et al., 2013; Cettolo et al., 2017; Akhbardeh et al., 2021), where annotators are shown both the source text and the translated target text. For DA, we again make use of the automatic re-segmentation as outlined in §3.1, which has several important benefits. First, it makes possible to choose only a subset of segments for manual annotation, which is often desirable for cost considerations while ensuring that the segmentation and choice of segments remain comparable across the evaluated systems. Second, it allows computing segment-level correlation with automatic metrics, thanks to consistent segmentation. Finally, it frees us from having to apply weighted averaging schemes when aggregating segment scores, in order to make up for the possibility of different systems using different segment granularity.

Figure 2 shows an example of how DA based on resegmented system outputs looks in practice. Note that the re-segmentation step described above may lead to cases where annotators are presented with segments of automatic translations that start or stop in the middle of a sentence (see example in Figure 2). Such apparent problems in the translation output may originate simply from the different segmentation but may be perfectly valid when considered in context. To avoid penalizing such situations, we provided translators not only with the source sentence and system translation but also with the system translation of the previous and following segments. The precise annotation instructions are given in the appendix. Providing more context to human annotators is additionally motivated by prior research demonstrating higher annotation quality by showing document context to annotators (Grundkiewicz et al., 2021).

Assessments were performed on a continuous scale between 0 and 100. Neither video nor audio

[9] Some initial work exists, for instance Bentivogli et al. (2021) do not find significant differences in preference, while Gaido et al. (2020) investigate speaker gender awareness as a concept that only direct models can learn in theory, since cascaded models lose this information through reliance on the intermediate transcription.

context was provided. Segments were shuffled and randomly assigned to annotators to avoid bias related to the presentation order. Annotations were conducted through a trusted vendor by a total of 77 professional translators (33 for English-German, 11 for English-Japanese, 33 for English-Chinese). Annotators were fluent in the source language and native in the target language.

DA scores were collected for all three considered language pairs, i.e. English-German, English-Japanese, and English-Chinese. For the TED domain, DA scores are collected over a subset of 1,000 randomly chosen segments. The same segments are chosen for all the systems.

3.3.2. MQM

As part of the evaluation campaign, a portion of the submissions were annotated using Multidimensional Quality Metrics (MQM; Lommel et al., 2014). MQM has been used in the WMT shared task series in recent years (Freitag et al., 2021) and is promising for detailed analyses of translation results. We include this data in our analysis in order to verify the quality of the collected DA scores. The MQM evaluation was focused only on English-to-Japanese simultaneous translation. See Agarwal et al. (2023) for more details.

3.3.3. Continuous Rating

In addition, the IWSLT 2023 evaluation campaign also included a human evaluation study using the *Continuous Rating* method (Javorský et al., 2022) for the English-to-German simultaneous translation within the TED domain. This evaluation method was selected as it has been specifically designed for the simultaneous translation regime. Again, we include this data in our analysis in order to verify the quality of the collected DA scores.

CR replicates the real-world translation scenario, in which the speaker receives incomplete translations while delivering the speech. Native German speakers fluent in English were assigned to listen to the source audio in English and continuously evaluate the quality of the German translation. The translation grew incrementally with respect to the timestamps corresponding to the source audio as the system generated each word. All systems followed the same reference segmentation of the source audio, but the human evaluation was on the talk level. For more details about the evaluation process, please refer to Agarwal et al. (2023) and Anastasopoulos et al. (2022).

4. Main Results

In our meta-evaluation, we compare human evaluation results with those computed with automatic metrics. For the latter, we use the full evaluation set (416 sentences for the ACL domain, and around 2,000 for the TED domain depending on language pair), while DA on TED was restricted to 1,000 segments. Following Agarwal et al. (2023), we used the new, more natural TED references unless otherwise noted.

We rely on computing correlations as our main analysis tool. One challenge is that the number of data points (systems) per evaluated condition was between only 2 and 20, which on the lower end was sometimes too small to obtain statistically significant results. For a more robust analysis, we therefore present both Pearson linear correlation (denoted ρ) and Spearman rank correlation (denoted r). Intuitively, ρ captures a metric's informativeness regarding the magnitude by which systems differ, while r measures its reliability for ranking of systems, both of which are generally of interest. All correlations are based on system-level scores. Statistically significant correlations at $p \leq .05$ are shown in bold font throughout.

An overview of the systems under evaluation is provided by Agarwal et al. (2023). Detailed data size statistics are in Table 2.

4.1. DA and Automatic Metrics

Table 3 shows the correlation between chrF and DA scores. We observe generally high ρ, while r tends to be lower in some cases. Note that no coefficients lower than 0.75 were statistically significant, suggesting an insufficient number of data points as the main reason for those very low correlation coefficients. The fact that results in many cases approach perfect correlation can be explained by observing that there was often a significant accuracy gap between the submitted systems (Agarwal et al., 2023). We speculate as another contributing factor that speech translation quality has not yet reached the level where automatic metrics become increasingly unreliable. This is in contrast to machine translation, where it is often found that for top-performing systems, automatic metrics have a poor correlation with human-produced assessments (Mathur et al., 2020).

Table 4 shows, among others, the correlation between COMET and DA scores. As can be observed in columns 4–7, COMET scores yielded higher correlations than chrF for all cases where both sides had statistically significant correlation coefficients. This indicates that COMET is robust enough to be applied to the speech translation scenario which includes additional challenges such as noisy sentence segmentation. However, prior work (Amrhein and Haddow, 2022, Appendix D1) casts some doubt on this conclusion, and a more thorough study of this issue is needed. Also, note that while the results show higher correlation for

Task	Language	Domain	Systems	Segments	Tokens
Offline	en-de	TED	7	7000	115633
Multilingual	en-de	ACL	3	1248	22128
Offline	en-de	ACL	7	2912	51632
Simultaneous	en-de	TED	5	5000	82595
Simultaneous	en-ja	TED	4	4000	66076
Offline	en-ja	TED	5	5000	82595
Offline	en-ja	ACL	5	2080	36880
Multilingual	en-ja	ACL	3	1248	22128
Offline	en-zh	TED	8	8000	132152
Offline	en-zh	ACL	8	3328	59008
Multilingual	en-zh	ACL	3	1248	22128
Simultaneous	en-zh	TED	2	2000	33038

Table 2: Data statistics for collected direct assessments: Number of systems, segments, and source side (reference transcript) tokens for each combination of task, language pair, and domain.

Task	Language	Domain	Systems	ρ		r	
Offline	en-de	TED	7	**0.99**	(p=0.00)	0.71	(p=0.07)
Multilingual	en-de	ACL	3	0.73	(p=0.48)	0.50	(p=0.67)
Offline	en-de	ACL	7	**0.81**	(p=0.03)	0.36	(p=0.43)
Simultaneous	en-de	TED	5	0.75	(p=0.15)	0.50	(p=0.39)
Simultaneous	en-ja	TED	4	**0.99**	(p=0.01)	0.20	(p=0.80)
Offline	en-ja	TED	5	**0.99**	(p=0.00)	**0.90**	(p=0.04)
Offline	en-ja	ACL	5	**0.99**	(p=0.00)	0.70	(p=0.19)
Multilingual	en-ja	ACL	3	0.95	(p=0.20)	0.50	(p=0.67)
Offline	en-zh	TED	8	**0.96**	(p=0.00)	**0.79**	(p=0.02)
Offline	en-zh	ACL	8	**0.75**	(p=0.03)	0.19	(p=0.65)
Multilingual	en-zh	ACL	3	−0.01	(p=0.99)	−0.50	(p=0.67)
Simultaneous	en-zh	TED	2	1.00	(p=1.00)	1.00	(p=1.00)

Table 3: Pearson correlation (ρ) and Spearman correlation (r) of DA scores vs. chrF scores.

COMET scores with DA than chrF with DA, the magnitude of this improvement remains somewhat unclear, because chrF correlations are generally already so high that there is not much room for improvement. Further experiments are needed to better understand whether the extent to which trainable metrics such as COMET outperform string comparison based metrics such as chrF is comparable to the large gap observed in text translation (Freitag et al., 2022).

To summarize: **DA is highly correlated with both automatic metrics in most but not all cases for our data, and COMET is found to outperform chrF in the speech translation scenario.**

4.2. MQM

The previous subsection investigates correlations between automatic metrics and human judgements in the form of DA scores, but we also wish to assess the soundness of the DA scores as a gold standard. While high correlation between DA and automatic metrics provides some positive indication, we now turn to MQM scores, available for portions of the English-Japanese data, as a more reliable means of verification of soundness of the DA scores. Table 5 shows correlations between the MQM score and other scores (DA, chrF, and COMET) at the system level. We observed clearly negative correlations[10] with all the scores. This is consistent with the findings above, and also further corroborates the robustness of the collected DA scores.

4.3. Continuous Rating

Continuing in this spirit, Table 6 compares the correlation of the Continuous Rating (CR) evaluation with the new Direct Assessment evaluations, and the two automatic metrics chrF and COMET on English-to-German simultaneous translations. The correlation between CR and DA of 0.95 demonstrates the validity of the DA scores. The correlation of the automatic metric COMET with DA is 0.94, which is relatively close to the correlation of 0.96 between COMET and DA (see Table 4). In-

[10] Note that an MQM score is a weighted sum of *error* scores.

Task	Lang.	Dom.	$\rho_{DA/\text{CHRF}}$	$r_{DA/\text{CHRF}}$	$\rho_{COMET/DA}$	$r_{COMET/DA}$	Original TED refs. $\rho_{DA/\text{CHRF}}$	Original TED refs. $r_{DA/\text{CHRF}}$
Offline	en-de	TED	**0.99**	0.71	**0.99**	0.68	**0.97**	0.61
Multi	en-de	ACL	0.98	**1.00**	**1.00**	**1.00**	–	–
Offline	en-de	ACL	**0.94**	**0.75**	**0.99**	**0.89**	–	–
Simul	en-de	TED	0.93	0.60	**0.96**	0.80	**0.97**	0.80
Simul	en-ja	TED	**0.99**	0.20	**1.00**	**1.00**	**0.99**	0.20
Offline	en-ja	TED	**0.99**	**0.90**	**1.00**	0.70	**0.99**	**0.90**
Offline	en-ja	ACL	**0.99**	0.60	**1.00**	0.70	–	–
Multi	en-ja	ACL	0.97	0.50	**1.00**	**1.00**	–	–
Offline	en-zh	TED	**0.97**	**0.89**	**1.00**	**0.96**	**0.98**	**0.89**
Offline	en-zh	ACL	**0.98**	**0.96**	**1.00**	**1.00**	–	–
Multi	en-zh	ACL	**1.00**	**1.00**	0.99	**1.00**	–	–
Simul	en-zh	TED	1.00	1.00	1.00	1.00	1.00	1.00

Table 4: Comparison of COMET and CHRF correlation with DA scores, respectively. In addition, the DA/CHRF correlation based on the original TED references are given.

Task	Language	Domain	Systems	$\rho_{DA/MQM}$	$\rho_{\text{CHRF}/MQM}$	$\rho_{\text{COMET}/MQM}$
Simultaneous	en-ja	TED	4	-0.97 (p=0.03)	-0.99 (p=0.01)	-0.98 (p=0.02)

Table 5: System-level correlation of DA, CHRF, and COMET against MQM for the 107 segments subset.

terestingly, the automatic metric CHRF seems to correlate with CR much less than CHRF with DA (0.75 vs. 0.93).

To summarize: **Comparison with both MQM and CR scores verifies the reliability of the DA scores, collected as described in Section 3.3.1.**

5. Further Analysis

5.1. New versus Old References

As indicated in Section 2.2.1, TED data included two kinds of references: The original references in subtitle style, and new references that were created for more naturalness. We now turn to the question of whether creating these new references helped make automatic evaluation more reliable. To this end, Table 4 shows correlations for the original references (rows 8–9), addition to the new references (rows 4–5). We find that correlation coefficients do not differ much between the two references. This indicates that the subtitle-style references are of adequate quality in the context of our data set, and that the effort to create more natural references is perhaps questionable. However, we stress that the situation might change if systems get more accurate, or if there are more systems that are very close in terms of accuracy. In both of these cases having high-quality references is expected to be beneficial for reliable evaluation.

To summarize: **In the context of our data, the original TED subtitles are equally suitable for evaluation as the new, natural TED references.**

5.2. Cross-Condition Correlations

The overlap between conditions such as task, domain, and languages in our data provides an opportunity to study the feasibility of cross-condition comparisons.

First, we combine *domains* by using the average CHRF score between ACL and TED test sets for the offline task. Such a comparison could be conducted to rank systems with regards to their cross-domain translation ability. It also provides an additional perspective to verify observations made in earlier sections. To this end, we observe in Table 7 that differences between original and new TED references exist but are minor, confirming the conclusions from Section 5.1. The Table also confirms COMET to be

Task	Language	Domain	Systems	$\rho_{DA/CR}$	$\rho_{\text{CHRF}/CR}$	$\rho_{COMET/CR}$
Simultaneous	en-de	TED	5	0.95 (p=0.01)	0.75 (p=0.15)	0.94 (p=0.02)

Table 6: System-level correlation of DA, CHRF, and COMET against Continuous Rating for English-to-German simultaneous translation data.

Task	Lang.	Dom.	$\rho_{DA/\text{CHRF}}$	$r_{DA/\text{CHRF}}$	$\rho_{DA/\text{COMET}}$	$r_{DA/\text{COMET}}$	Original TED refs. $\rho_{DA/\text{CHRF}}$	Original TED refs. $r_{DA/\text{CHRF}}$
Offline	en-de	Avg	**0.98**	**0.75**	**0.99**	**0.89**	**0.99**	**0.75**
Offline	en-ja	Avg	**0.99**	0.70	**1.00**	**0.90**	**0.99**	0.70
Offline	en-zh	Avg	**0.98**	**0.86**	**1.00**	**1.00**	**0.98**	**0.96**

Table 7: Correlation of DA and CHRF after *averaging* scores from TED and ACL domains.

Lang.	Dom.	Sys.	$\rho_{DA/\text{CHRF}}$	$r_{DA/\text{CHRF}}$	$\rho_{DA/\text{COMET}}$	$r_{DA/\text{COMET}}$	Original TED references $\rho_{DA/\text{CHRF}}$	Original TED references $r_{DA/\text{CHRF}}$
en-de	TED	11	**0.97**	**0.88**	**0.97**	**0.85**	**0.96**	**0.86**
en-de	ACL	10	**0.92**	**0.70**	**0.98**	**0.88**	–	–
en-ja	TED	9	**0.96**	**0.75**	**0.97**	0.57	**0.96**	**0.75**
en-ja	ACL	8	**0.93**	0.55	**0.99**	**0.86**	–	–
en-zh	TED	9	**0.91**	**0.83**	**0.99**	**0.93**	**0.91**	**0.75**
en-zh	ACL	10	**0.97**	**0.98**	**0.99**	**1.00**	–	–

Table 8: Correlations of DA and CHRF on original/new references, computed over systems across task types (offline+simultaneous for TED, offline+multilingual for ACL).

better correlated with human judgments than CHRF, similar to findings in Section 4.1.

Another question we might pose is whether automatic metrics can be compared across different conditions. In speech translation, especially the case of comparing across task types (offline, multilingual, simultaneous) is of practical relevance, e.g. where one might wish to judge how well a given offline system is perceived by users in comparison to a given simultaneous system. Practitioners might also wish to compare systems across domains or even languages, for example if through experience a threshold on an automatic metric has been determined above which user experience is satisfactory, and one wonders whether the same threshold can be applied to a different domain or target language. Factors that would hinder such a comparison include any phenomenon that systematically biases an automatic score in comparison to human judgement, such as morphological complexity (when comparing across target languages), translation artifacts introduced by simultaneous translation systems (when comparing task types), or peculiarities of the domain that have an overly strong or weak influence on the metric. While researchers often refrain from such comparisons due to potential issues along these lines, our data, due to its overlapping nature, provides some opportunity to investigate the validity of such comparisons.

We conduct this analysis by simply computing

Task	Language	$\rho_{DA/\text{CHRF}}$	$r_{DA/\text{CHRF}}$	$\rho_{DA/\text{COMET}}$	$r_{DA/\text{COMET}}$	Original TED refs. $\rho_{DA/\text{CHRF}}$	Original TED refs. $r_{DA/\text{CHRF}}$
Offline	en-de	**0.74**	**0.53**	−0.30	−0.01	**0.66**	0.31
Offline	en-ja	**0.76**	**0.64**	0.00	0.39	**0.84**	**0.64**
Offline	en-zh	**0.90**	**0.74**	−0.36	−0.03	0.34	0.20

Table 9: Correlation of DA scores and automatic metrics, computed over systems across both data domains. We show only the offline task, the only task that includes both domains.

Task	Domain	$\rho_{DA/\text{CHRF}}$	$r_{DA/\text{CHRF}}$	$\rho_{DA/\text{COMET}}$	$r_{DA/\text{COMET}}$	Original TED refs. $\rho_{DA/\text{CHRF}}$	Original TED refs. $r_{DA/\text{CHRF}}$
Offline	TED	**0.87**	**0.94**	**0.70**	**0.52**	**0.67**	**0.67**
Multi	ACL	**0.67**	0.60	0.58	0.49	–	–
Offline	ACL	**0.78**	**0.78**	**0.65**	**0.45**	–	–
Simul	TED	**0.88**	**0.93**	**0.69**	0.01	**0.79**	**0.82**

Table 10: Correlation of DA scores and CHRF, computed over systems across all three language pairs.

the correlation between DA and automatic metric jointly across systems from the different conditions. First, we observe that jointly considering systems from different *task conditions* (offline, multilingual, simultaneous) yields high correlations (Table 8). As discussed, this is desirable in speech translation with its multitude of task types.

Next, doing the same analysis across domains (Table 9) and target languages (Table 10) reveals some interesting findings: Correlations in rows 3–4 (chrF with new TED references) are significantly lower in comparison but still strongly positive, indicating that these comparisons, with some care, can be informative. In contrast, chrF against subtitle references, and COMET, are more poorly correlated. This indicates that more accurate TED references are of value under such challenging conditions, and that COMET is more brittle than chrF.

To summarize: **We find further evidence for observations from previous sections, and that it is safe to compare systems across the common speech translation task types such as offline and simultaneous systems.**

6. Conclusion

We have presented a first step towards establishing human evaluation practices in speech translation. Our approach includes direct assessment based on automatically segmented inputs. Comparison with MQM and CR shows the robustness of our approach. Correlations with automatic metrics are generally high, and (in line with findings in text translation) COMET is observed to slightly outperform chrF, but more experiments are needed to corroborate these findings. Future work may investigate whether evaluation in which human annotators are presented with audio segments rather than text segments could be a feasible improvement.

7. Bibliographical References

Milind Agarwal, Sweta Agrawal, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Mingda Chen, William Chen, Khalid Choukri, Alexandra Chronopoulou, Anna Currey, Thierry Declerck, Qianqian Dong, Kevin Duh, Yannick Estève, Marcello Federico, Souhir Gahbiche, Barry Haddow, Benjamin Hsu, Phu Mon Htut, Hirofumi Inaguma, Dávid Javorský, John Judge, Yasumasa Kano, Tom Ko, Rishu Kumar, Pengwei Li, Xutai Ma, Prashant Mathur, Evgeny Matusov, Paul McNamee, John P. McCrae, Kenton Murray, Maria Nadejde, Satoshi Nakamura, Matteo Negri, Ha Nguyen, Jan Niehues, Xing Niu, Atul Kr. Ojha, John E. Ortega, Proyag Pal, Juan Pino, Lonneke van der Plas, Peter Polák, Elijah Rippeth, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Yun Tang, Brian Thompson, Kevin Tran, Marco Turchi, Alex Waibel, Mingxuan Wang, Shinji Watanabe, and Rodolfo Zevallos. 2023. Findings of the IWSLT 2023 Evaluation Campaign. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 1–61, Toronto, Canada (in-person and online). Association for Computational Linguistics.

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.

Chantal Amrhein and Barry Haddow. 2022. Don't discard fixed-window audio segmentation in speech-to-text translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 203–219, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Antonios Anastasopoulos, Loïc Barrault, Luisa Bentivogli, Marcely Zanon Boito, Ondřej Bojar, Roldano Cattoni, Anna Currey, Georgiana Dinu, Kevin Duh, Maha Elbayad, Clara Emmanuel, Yannick Estève, Marcello Federico, Christian Federmann, Souhir Gahbiche, Hongyu Gong, Roman Grundkiewicz, Barry Haddow, Benjamin Hsu, Dávid Javorský, Věra Kloudová, Surafel Lakew, Xutai Ma, Prashant Mathur, Paul McNamee, Kenton Murray, Maria Nădejde, Satoshi Nakamura, Matteo Negri, Jan Niehues, Xing Niu, John Ortega, Juan Pino, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Yogesh Virkar, Alexander Waibel, Changhan Wang, and Shinji Watanabe. 2022. Findings of the IWSLT 2022 evaluation campaign. In *Proceedings of the 19th*

International Conference on Spoken Language Translation (IWSLT 2022), pages 98–157, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Luisa Bentivogli, Mauro Cettolo, Marco Gaido, Alina Karakanta, Alberto Martinelli, Matteo Negri, and Marco Turchi. 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 and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2873–2887, Online. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

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.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021. 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.

Marco Gaido, Beatrice Savoldi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2020. Breeding gender-aware direct speech translation systems. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3951–3964, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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.

Roman Grundkiewicz, Marcin Junczys-Dowmunt, Christian Federmann, and Tom Kocmi. 2021. On user interfaces for large-scale document-level human evaluation of machine translation outputs. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 97–106, Online. Association for Computational Linguistics.

Dávid Javorský, Dominik Macháček, and Ondřej Bojar. 2022. Continuous rating as reliable human evaluation of simultaneous speech translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 154–164, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine

translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional Quality Metrics (MQM): A Framework for Declaring and DescribingTranslation Quality Metrics. *Revista Tradumàtica: tecnologies de la traducció*, 12:455–463.

Mingbo Ma, Liang Huang, Hao Xiong, Renjie Zheng, Kaibo Liu, Baigong Zheng, Chuanqiang Zhang, Zhongjun He, Hairong Liu, Xing Li, Hua Wu, and Haifeng Wang. 2019. STACL: Simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3025–3036, Florence, Italy. Association for Computational Linguistics.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Evgeny Matusov, Gregor Leusch, Oliver Bender, and Hermann Ney. 2005. Evaluating machine translation output with automatic sentence segmentation. In *Proceedings of the Second International Workshop on Spoken Language Translation*, Pittsburgh, Pennsylvania, USA.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Elizabeth Salesky, Kareem Darwish, Mohamed Al-Badrashiny, Mona Diab, and Jan Niehues. 2023. Evaluating multilingual speech translation under realistic conditions with resegmentation and terminology. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 62–78, Toronto, Canada (in-person and online). Association for Computational Linguistics.

A. Appendix: Annotator instructions

in order to minimize the impact of segmentation issues during direct assessment (DA), the following instructions were given to human annotators:
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 judgment.