MT Metrics Correlate with Human Ratings of Simultaneous Speech Translation (Technical Report)

Dominik Macháček[1] and **Ondřej Bojar**[1] and **Raj Dabre**[2]

Charles University, Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics[1]

National Institute of Information and Communications Technology, Kyoto, Japan[2]
[1]{machacek,bojar}@ufal.mff.cuni.cz, [2]raj.dabre@nict.go.jp

Abstract

There have been several studies on the correlation between human ratings and metrics such as BLEU, chrF2 and COMET in machine translation. Most, if not all consider full-sentence translation. It is unclear whether human ratings of simultaneous speech translation Continuous Rating (CR) correlate with these metrics or not. Therefore, we conduct an extensive correlation analysis of CR and the aforementioned automatic metrics on evaluations of candidate systems at English-German simultaneous speech translation task at IWSLT 2022. Our studies reveal that the offline MT metrics correlate with CR and can be reliably used for evaluating machine translation in the simultaneous mode, with some limitations on the test set size. This implies that automatic metrics can be used as proxies for CR, thereby alleviating the need for human evaluation.

1 Introduction

The current approach to evaluate SST systems that have text as the output modality is to use automatic metrics which are designed for offline text-to-text machine translation (MT), alongside to other measures for latency and stability. The most used metric, according to the meta-evaluation of general MT research (Marie et al., 2021), is BLEU (Papineni et al., 2002), however, other state-of-the-art metrics such as chrF2 (Popović, 2017) and COMET (Rei et al., 2020) are shown to correlate with human judgements more than BLEU (Freitag et al., 2021). Researchers tend to use these offline automatic metrics in simultaneous speech translation despite the fact that there is no explicit evidence that they correlate with human ratings.

However, simultaneous speech-to-text translation has different characteristics than offline text-to-text MT. For example, when the users are following subtitles in real-time, they have limited time for reading and comprehension as they can not fully control the reading pace by themselves. Therefore, they may be less sensitive to subtle grammar and factual flaws than while reading a text document without any time constraints. They may also prefer brevity and simplicity over verbatim word-for-word translation. Even if the reference is brief and simpler than the original, there may be lots of variants that the BLEU score and other MT metrics may not value as correct.

Furthermore, SST and MT differ in their input modalities. MT sources are assumed to originate as texts, while the SST source is a speech given in certain situation, accompanied by para-linguistic means and specific knowledge (context) shared by the speaker and listener. Transcribing speech to text for use in the offline evaluation of SST may be limiting. The human evaluation of SST should therefore reflect the simultaneity of original video or audio.

In this paper, we aim to establish whether the usage of these automatic metrics is an appropriate way of evaluating SST. To do so, we calculate correlations between human judgements and BLEU as well as other MT metrics in simultaneous mode. To this end, we analyze the results of the simultaneous speech translation task from English to German at IWSLT 2022 (Anastasopoulos et al., 2022). In this task, there are 5 competing systems and human interpreting that are manually rated by bilingual judges in a simulated real-time event. Our studies show that BLEU does indeed correlate with human judgements of simultaneous translations under the same conditions as in offline text-to-text MT: on substantially large number of sentences or references. Furthermore, chrF2 and COMET exhibit similar correlations. To the best of our knowledge, we are the first to explicitly establish the correlation between automatic offline metrics with human SST ratings, indicating that they may be safely used in SST evaluation.

2 Data of Human Ratings

2.1 IWSLT22 En-De Simultaneous Translation Task

In IWSLT 2022 (Salesky et al., 2022), there were multiple tasks and language pair tracks, as described in “Findings” (Anastasopoulos et al., 2022), however, we focus only on the English-to-German Simultaneous Translation Task because it is the only one that was also evaluated manually in simultaneous mode. The task focused on speech-to-text translation and was reduced to translation of individual sentences. The segmentation of the source audio to sentences was provided by organizers, and not by the systems themselves. The source sentence segmentation that was used in human evaluation was gold (oracle). It only approximates the realistic setup where the segmentation would be provided by an automatic system, e.g. Tsiamas et al. (2022), and may be partially incorrect and cause more translation errors than the gold segmentation.

The simultaneous mode in Simultaneous Translation Task means that the source is provided gradually, one audio chunk at a time. After receiving each chunk, the system decides to either wait for more source context, or produce target tokens. Once the target tokens are generated, they can not be rewritten.

The participating systems are submitted and studied in three latency regimes: low, medium and high. It means that the maximum Average Lagging (Ma et al., 2019) between the source and target on validation set must be 1, 2 or 4 seconds in “computationally unaware” simulation where the time spent by computation, and not by waiting for context, is not counted. One system in low latency did not pass the latency constraints (see Findings, page 44, numbered 141), but it is evaluated manually anyway.

Computational unaware latency was one of the main criteria in IWSLT 2022. It means that the participants did not need to focus on low latency implementation, as it is more a technical and hardware issue than a research task. However, the subtitle timing in manual evaluation was created in a way that waiting for the first target token is dropped, and then it continues with computationally aware latency.

2.2 Highlighting Findings

The Findings of IWSLT22 (Anastasopoulos et al., 2022) are available in PDF. The most up-to-date version (version 2) is 61 pages long[1]. We highlight the relevant parts of Findings with page numbers in Table 1 so that we can refer to them easily.

Note that findings are a part of the conference proceedings (Salesky et al., 2022) as a chapter in a book. The order of findings pages in PDF does not match the page numbers at the footers.

Also note that in Section 2.4 on page 4 (in PDF, 101 in Proceedings), there is a description of MLLP-VRAIN and that corresponds to the system denoted as UPV in all other tables and figures.

2.3 Continuous Rating (CR)

Continuous Rating (CR, Javorský et al., 2022; Macháček and Bojar, 2020) is a method for human assessment of SST quality in a simulated online event. An evaluator with knowledge of the source and target language watches a video (or listens to an audio) document with subtitles created by the SST system which is being evaluated. The evaluator is asked to continuously rate the quality of the translation by pressing buttons with values 1 (the worst) to 4 (the best). Each evaluator can see every document only once, to ensure one-pass access to the documents, as in a realistic setup.

CR is analogous to Direct Assessment (Graham et al., 2015), which is a method of human text-to-text MT evaluation in which a bilingual evaluator expresses the MT quality by a number on a scale. It is natural that individual evaluators have different opinions, and thus it is a common practice to have multiple evaluators evaluate the same outputs and then report the mean and standard deviation of evaluation scores, or the results of statistical significance tests that compare the pairs of candidate systems and show how confident the results are.

Javorský et al. (2022) showed that CR relates to comprehension of foreign language documents by SST users. Using CR alleviates the need to evaluate comprehension by factual questionnaires that are difficult to prepare, collect and evaluate. Furthermore, Javorský et al. (2022) show that bilingual evaluators are reliable.

Criteria of CR In IWSLT 2022, the evaluators were instructed that the primary criterion in CR should be meaning preservation (or adequacy), and other aspects such fluency should be secondary. The instructions do not mention readability due to output segmentation frequency or verbalizing

[1] https://aclanthology.org/2022.iwslt-1.10v2.pdf

marker	PDF page	numbered page	description
Section 2	3-5	100-102	Simultaneous Speech Translation Task
Figure 1	6	103	Quality-latency trade-off curves
Section 2.6.1	5	102	Description of human evaluation
Figure 5	8	105	Manual scores vs BLEU (plot)
Two Test Sets (paragraph)	39	136	Non-native subset
Test data (paragraph)	9	106	Common (native) subset of test data
Automatic Evaluation Results	44	141	Latency and BLEU results (table)
A1.1 (appendix)	38-39	135-136	Details on human evaluation
Table 17	48	145	Test subsets duration
Table 18	48	145	Manual scores and BLEU (table)

Table 1: Relevant parts of IWSLT22 Findings (`https://aclanthology.org/2022.iwslt-1.10v2.pdf`) for En-De Simultaneous Speech Translation task and human evaluation.

non-linguistic sounds such as “laughter”, despite that the system candidates differ in these aspects. Unfortunately there is also no way to accurately detect, to which extent the evaluators followed said criteria.

2.4 Candidate Systems

Automatic SST systems There are 5 evaluated SST systems: FBK (Gaido et al., 2022), NAIST (Fukuda et al., 2022), UPV (Iranzo-Sánchez et al., 2022), HW-TSC (Wang et al., 2022), and CUNI-KIT (Polák et al., 2022). More details are in system description papers. They are also summarized in Findings in Section 2.4.

Human Interpreting In order to compare the state-of-the-art SST with human reference, the organizers hired one expert human interpreter to simultaneously interpret all the test documents. Then, they employed annotators to transcribe the voice into texts. The annotators worked in offline mode. The transcripts were then formed as subtitles and were used in CR evaluation the same way as SST.

However, human interpreters use their own segmentation to translation units so that they often do not translate one source sentence as one target sentence. There is no alignment of the source sentences to interpreting chunks. We can not use the automatic metrics that rely on the same sentence segmentation of the candidate and reference (e.g. COMET) for interpreting, or they must be first adjusted, e.g. calculating BLEU and chrF2 on the whole documents instead of on aligned sentences. In this analysis, we therefore use interpreting only in Section 3, but not for correlating the MT metrics in Section 4.

2.5 Evaluation Data

There are two subsets of evaluation data used in IWSLT22 En-De Simultaneous Translation task. The “Common” subset consists of TED talks. See the description in Findings on page 9 (numbered as 106). The speakers in TED talks are native. “Non-Native” subset consists of mock business presentations of European high school students (Macháček et al., 2019), and of presentations by representatives of European supreme audit institutions. This subset is described in Findings on page 39 (numbered page 136). The duration statistics of audio documents in both test sets are in Findings in Table 17 on page 48 (numbered 145).

3 Aggregating Continuous Ratings

In this section, we question the interpretation of Continuous Rating that has an impact on aggregation of the individual clicks of the rating buttons to the final score of the whole document or set of documents.

We found two definitions that can yield different results in certain situations: (1) The rating is valid in an instant time point when the evaluator clicked the rating button. The final score is the average of all clicks, each click has the equal weight. We denote it as CR.

(2) The rating is assigned to the time interval between the clicks, or between the last click and the end of the document. The length of the interval is considered in averaging. The final score is the average of ratings weighted by interval lengths when the rating is valid. We denote it as CRi.

To express them rigorously, let us have a document of duration T, and n ratings (r_i, t_i), where $i \in \{1, \ldots, n\}$ is an index, $r_i \in \{1, \ldots, 4\}$ is the

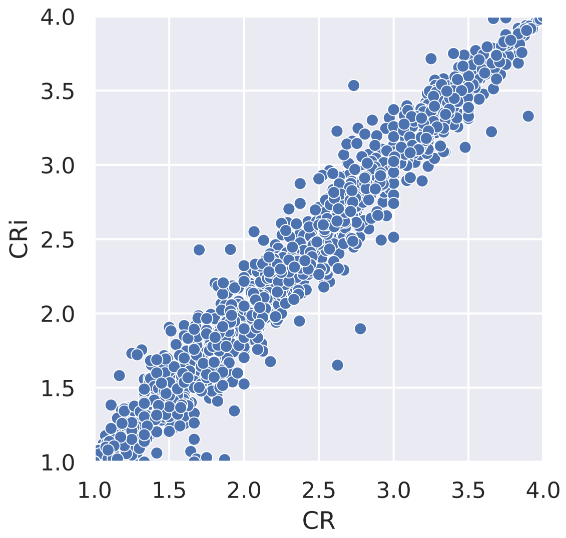

Figure 1: Relation between weighted interval averaging of continuous rating (CRi, y-axis) and average of all ratings (CR, x-axis) for each annotation of each document (blue data points).

rated value and $0 \leq t_1 < \cdots < t_n \leq T$ are times when the ratings were recorded.

Then, the definitions are as follows:

$$CR = \frac{1}{n} \sum_{i=1}^{n} r_i$$

$$CRi = \frac{1}{T - t_1} \left(\sum_{i=1}^{n-1} (t_{i+1} - t_i) r_i + (T - t_n) r_n \right)$$

If the judges press the rating buttons regularly, with a uniform frequency, then both definitions give equal scores. Otherwise, the CR and CRi may differ and may yield even opposite conclusions. For example, pressing "1" twelve times in one minute, then "4" and then waiting for one minute results in different scores: $CR = 1.2$, $CRi = 2$.

To examine the relationship between these definitions, we counted CR and CRi for each annotation of each document in the evaluation campaign. The results are in Figure 1 where we observe correlation between the two definitions. The Pearson correlation coefficient is 0.98, which indicates a strong correlation.

Summary Based on the correlation scores we observed, we conclude that both definitions are interchangeable, and any of them can be used in further analysis.

Averaged document ratings

subsets	num.	BLEU	chrF2	COMET
both	823	0.65	0.73	0.80
Common	228	0.42	0.63	0.76
Non-native	595	0.70	0.70	0.75

All document ratings

subsets	num.	BLEU	chrF2	COMET
both	1584	0.61	0.68	0.73
Common	441	0.37	0.57	0.68
Non-native	1143	0.64	0.64	0.67

Table 2: Pearson correlation coefficients for CR vs MT metrics BLEU, chrF2 and COMET for averaged document ratings by all 5 SST systems and 3 latency regimes (upper), and all ratings (lower). When the coefficient is less than 0.6 (in gray), the correlation is not considered as strong. Significance values are $p < 0.01$ in all cases, meaning strong confidence.

4 Correlation of CR and MT Metrics

In this section, we study the correlation of CR and MT metrics BLEU, chrF2 and COMET. We measure it on the level of documents, and not on the test set level, because we have substantially large data for significant results. There are 60 evaluated documents (17 in the Common subset and 43 in Non-native) and 15 system candidates (5 systems, each in 3 latency regimes), which yields 900 data points. There are only 15 in test set level, or 30 in subset level.

We discovered that CUNI-KIT system outputs are tokenized, while the others are detokenized. Therefore, we first detokenized CUNI-KIT outputs. Then, we removed the final end of sequence token (`</s>`) from the outputs of all systems. Finally, we calculated BLEU and chrF2 using sacreBLEU[2] (Post, 2018), and COMET (Rei et al., 2020) with `wmt20-comet-da` model.

In total, there are 1584 rating sessions of 900 candidate document translations with approximately 1.76 rating sessions per candidate document output. They differ by the evaluator. Some rating sessions were recorded, but excluded from further processing due to insufficient number of rating button clicks, see paragraph "Processing of Collected Rankings" in Findings on page 39, numbered 139.

We aggregate the individual rating clicks in each session by plain average (CR definition in Sec-

[2] Metric signatures:
BLEU|nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.1,
chrF2|nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.1

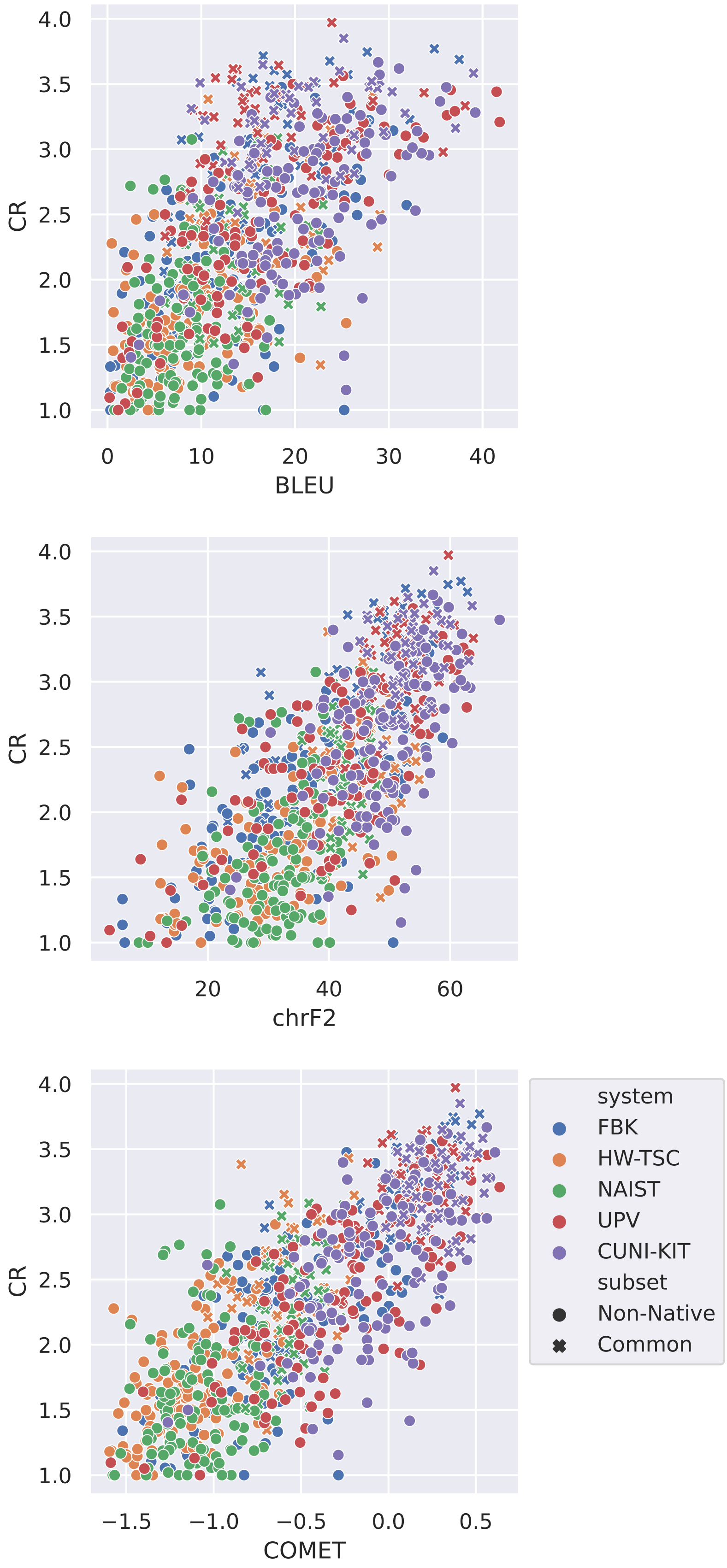

Figure 2: Averaged document CR vs MT metrics BLEU, chrF2 and COMET on both subsets.

tion 3) to get the CR scores. Then, we average the CR of the same documents and candidate translations, and we show them in Figure 2. In Table 2, we report the correlation coefficients with averaging and without, together with the number of observations.

Pearson correlation is considered as strong if the coefficient is larger than 0.6. The results show strong correlation (above 0.65 Pearson correlation coefficient) of CR with BLEU, chrF2 and COMET in document level on both test subsets. When we consider only one subset, the correlation is lower, but still strong for chrF2 and COMET (0.63 and 0.76). It is because the Common subset is generally translated better than Non-Native, so with only one subset, there are data points on a smaller span of the axis and there is a larger proportion of outliers.

It is not a case of BLEU on the Common subset with Pearson correlation coefficient 0.42. We assume it is because BLEU is designed for the use on a substantially large test set, but we use it on short single documents. However, BLEU strongly correlates with chrF2 and COMET (0.81 and 0.62 on the Common subset). BLEU also correlates with CR on the level of test sets, as reported in Findings in the caption of Table 18 (page 48, numbered 145).

Summary Based on the correlation results above, we conclude that BLEU, chrF2 and COMET can be used for reliable assessment of human judgement of SST quality at least on the level of test sets. chrF2 and COMET are also reliable at the document level.

Discussion of Limitations Let us remark that our analysis has limitations. The data that we analyzed are limited to only one English-German language pair, 5 SST systems from IWSLT 2022 and three domains. All the systems were supervised on translations. They do not aim to mimic interpretation with shortening, summarization or redundancy reduction, and they do not use document context. The MT metrics are good for evaluating individual sentence translations and that is important, but not the only subtask of SST. We assume that some future systems created with a different approach may show divergence of CR and the offline MT metrics.

5 Conclusion

In this technical report, we analyzed results of English-German Simultaneous Translation Task. We compared two interpretations of CR and aggregations of CR button clicks for document-level, and we showed that they strongly correlate and are interchangeable. Next, we discovered that CR correlates with BLEU, chrF2 and COMET of the system candidates and can be reliably used in simultaneous machine translation development if the test set size is large.

Acknowledgements

We are thankful to Dávid Javorský and Peter Polák for their reviews.

We will be very thankful to all who share this report with all members of the research community by word of mouth and by social media, as we are not present on the most relevant ones except e-mail and we do not publish it through a conference.

This research was partially supported by the grants 19-26934X (NEUREM3) of the Czech Science Foundation, “Grant Schemes at CU” (reg. no. CZ.02.2.69/0.0/0.0/19_073/0016935), SVV project number 260 575, and 398120 of the Grant Agency of Charles University.

References

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.

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.

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 2022)*, pages 286–292, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

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 2022)*, pages 177–189, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, and Nitika Mathur. 2015. Accurate evaluation of segment-level machine translation metrics. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1183–1191, Denver, Colorado. 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-Cerdà, 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 2022)*, pages 255–264, Dublin, Ireland (in-person and 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*, Stroudsburg, PA, USA. Association for Computational Linguistics, Association for Computational Linguistics.

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.

Dominik Macháček and Ondřej Bojar. 2020. Presenting simultaneous translation in limited space. In *Proceedings of the 20th Conference Information Technologies - Applications and Theory (ITAT 2020), Hotel Tyrapol, Oravská Lesná, Slovakia, September 18-22, 2020*, volume 2718 of *CEUR Workshop Proceedings*, pages 34–39. CEUR-WS.org.

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. Springer International Publishing.

Benjamin Marie, Atsushi Fujita, and Raphael Rubino. 2021. Scientific credibility of machine translation research: A meta-evaluation of 769 papers. 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 7297–7306, Online. Association for Computational Linguistics.

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.

Peter Polák, Ngoc-Quan Pham, Tuan Nam Nguyen, Danni Liu, Carlos Mullov, Jan Niehues, Ondřej Bojar, and Alexander Waibel. 2022. CUNI-KIT system for simultaneous speech translation task at IWSLT 2022. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 277–285, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. 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, Brussels, Belgium. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Unbabel's participation in the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 911–920, Online. Association for Computational Linguistics.

Elizabeth Salesky, Marcello Federico, and Marta Costa-jussà, editors. 2022. *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*. Association for Computational Linguistics, Dublin, Ireland (in-person and online).

Ioannis Tsiamas, Gerard I. Gállego, José A. R. Fonollosa, and Marta R. Costa-jussà. 2022. SHAS: Approaching optimal Segmentation for End-to-End Speech Translation. In *Proc. Interspeech 2022*, pages 106–110.

Minghan Wang, Jiaxin Guo, Yinglu Li, Xiaosong Qiao, Yuxia Wang, Zongyao Li, Chang Su, Yimeng Chen, Min Zhang, Shimin Tao, 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 2022)*, pages 247–254, Dublin, Ireland (in-person and online). Association for Computational Linguistics.