Evaluating Automatic Subtitling: Correlating Post-editing Effort and Automatic Metrics

Alina Karakanta[1*], Mauro Cettolo[2], Matteo Negri[2], Luisa Bentivogli[2]
[1]Leiden University Centre for Linguistics, Leiden, The Netherlands
[2]Fondazione Bruno Kessler, Trento, Italy
a.karakanta@hum.leidenuniv.nl {cettolo,negri,bentivo}@fbk.eu

Abstract

Systems that automatically generate subtitles from video are gradually entering subtitling workflows, both for supporting subtitlers and for accessibility purposes. Even though robust metrics are essential for evaluating the quality of automatically-generated subtitles and for estimating potential productivity gains, there is limited research on whether existing metrics, some of which directly borrowed from machine translation (MT) evaluation, can fulfil such purposes. This paper investigates how well such MT metrics correlate with measures of post-editing (PE) effort in automatic subtitling. To this aim, we collect and publicly release a new corpus containing product-, process- and participant-based data from post-editing automatic subtitles in two language pairs (en→de,it). We find that different types of metrics correlate with different aspects of PE effort. Specifically, edit distance metrics have high correlation with technical and temporal effort, while neural metrics correlate well with PE speed.

Keywords: automatic subtitling, post-editing, machine translation, evaluation, productivity

1. Introduction

Automatic subtitling is becoming a task of increasing interest for the MT community, practitioners and the audiovisual industry. High-quality automatic subtitling systems have the potential to increase subtitlers' productivity and provide access to audiovisual products for persons facing linguistic and sensory barriers when professional subtitling is not available. Automatic subtitling includes automatic translation of the speech, timestamp prediction (spotting) and segmentation of the text into subtitles. Each of these components has an effect on both quality and the post-editing (PE) process (Carroll and Ivarsson, 1998). However, the way subtitlers interact with automatically generated subtitles has not been yet explored. Previous studies in PE for subtitling (Volk et al., 2010; de Sousa et al., 2011; Etchegoyhen et al., 2014; Matusov et al., 2019; Koponen et al., 2020a; Huang and Wang, 2023) have assessed quality and productivity based on translation edits, without thoroughly studying the effort of adjusting the timestamps and segmentation, which is an integral part of the subtitling process. In addition, several automatic metrics have been proposed for assessing the quality of automatically-generated subtitles (Cherry et al., 2021; Karakanta et al., 2022; Wilken et al., 2022), while new neural-based metrics are coming with the promise of more robust evaluation even for challenging domains (Rei et al., 2020; Sellam et al., 2020; Zhang et al., 2020). However, to date there exists no study on the usefulness of automatic metrics for assessing productivity and quality in automatic subtitling. Such studies are vital for guiding researchers, developers and users in identifying which automatic metrics are most predictive of human performance and perceived quality in real-world automatic subtitling applications.

We fill this gap by performing a correlation analysis of automatic MT metrics with human measures of post-editing effort in automatic subtitling. To this aim, we collect and publicly release PE subtitling data in a real use case scenario where three professional subtitlers edit automatically generated and spotted subtitles in two language pairs (en→de,it).

Our contributions can be summarised as:

- We release a new data set, containing product (subtitles), process (time, keystrokes) and participant-based data in subtitling PE.
- By analysing the correlation between automatic MT metrics and PE effort, both at subtitle and task level, we show that edit distance metrics are good estimators for technical and temporal effort, while neural metrics also correlate well with PE speed.

The corpus and related documentation is publicly available through the CLARIN infrastructure.[1] More details on the project can be found at https://mt.fbk.eu/must-cinema-pe/.

2. Data Collection

2.1. Dataset

The data edited by the subtitlers comes from the MuST-Cinema corpus[2] (Karakanta et al., 2020), a speech subtitling corpus, compiled from subtitles

* Work done when in Fondazione Bruno Kessler.

[1]http://hdl.handle.net/10032/tm-a2-y2
[2]https://mt.fbk.eu/must-cinema/

LREC-COLING 2024, pages 6363–6369

of TED Talks.The MuST-Cinema test set contains 9 English single-speaker talks, amounting to 1 hour of video (545 sentences/10k words). The data collection was performed for English→{Italian,German}.

2.2. System and tool

Post-editing was performed in a novel automatic subtitling tool, Matesub.[3] Matesub features automatic transcription and translation, automatic generation of timestamps for the translated subtitles – a process called automatic spotting (or autospotting)– and automatic segmentation of the translated audio into subtitles. The process is completely automatic, and thus does not rely on a human source version of the subtitles (template). The subtitlers are presented with the video on which the subtitles appear. Subtitle blocks are shown as boxes along the timeline at the bottom of the screen. The position and length (duration) of the boxes can be adjusted using the mouse to match the beginning and end of the spoken utterance and to accommodate the time the subtitle should remain on screen. Matesub contains all the functionalities of typical subtitling editors, thus it is representative of subtitlers' real working settings.

2.3. Task

To guarantee high quality data, we relied on professional subtitlers with experience in subtitling and MTPE (minimum 2 years of experience on each) and regular users of Matesub, who were hired through a language service provider. Following *ad hoc* post-editing guidelines, and after a preliminary test session, one subtitler post-edited the German output and two subtitlers the Italian output.

Post-editing was conducted in 4 sessions, each one containing 3 tasks. The nine TED videos were split into 12 tasks (corresponding to entire talks or parts of them in case of longer talks), so that the average duration of the video for each task is 4 minutes. To avoid fatigue effects, the subtitlers took a break of at least 15 minutes between tasks. To reduce the possible influence of learning effects, the order of the talks was reversed for the second Italian subtitler.

The subtitle files (.srt) produced by the tool (original) were downloaded before the beginning of the task, while the final (PE) .srt files in the target language were downloaded after its completion. Per-subtitle data were recorded in process logs implemented in Matesub. An example of a process log can be seen in Table 1. The log records the original (automatic) and final (post-edited) subtitle text, original-final timestamps and time activity (time spent on each subtitle), along with other metadata, such as task and subtitle id. Since to date there exists no subtitling tool recording keystrokes, keystrokes were logged externally with InputLog (Leijten and Waes, 2013). Additionally, the subtitlers recorded their screen while post-editing, which helps investigate their editing decisions and identify outliers. At the end, the subtitlers completed a questionnaire to collect feedback on their user experience and perceptions of automatic subtitling, as per Koponen et al. (2020b).

[3] https://matesub.com/

2.4. Corpus structure

For each of the language pairs (en→de/it), the structure of the collected data is as follows:

- **srt**: subtitle files for each of the 9 TED videos for the system output (original) and each post-editor (it1, it2).
- **Keystroke logs**: one log file for each of the 12 tasks as .csv.
- **Process logs**: one log per subtitler containing all 12 tasks as .csv.
- **Parallel data**: the collected data as additional MuST-Cinema references (aligned at sentence-level).

The parallel data directory contains *i)* the presegmented .wav files of the MuST-Cinema release, *ii)* the aligned .txt files consisting of the source transcription (src), the target translation (ref) and the .yaml files also from MuST-Cinema, *iii)* the subtitling system output (sys), and *iv)* the PE versions by each subtitler (PE1, PE2). The system subtitles and the PE versions were manually aligned at the sentence level with the MuST-Cinema reference to ensure maximum quality. As in MuST-Cinema, the subtitle boundaries inside the sentences were marked with the symbols `<eob>` and `<eol>`. For en→it, we collected 1,199 subtitles for PE1 and 1,208 subtitles for PE2, while for en→de 1,198 subtitles. These correspond to 545 sentences for Italian and 542 sentences for German.

3. Experimental setup

Based on the collected data, we perform a correlation analysis of automatic metrics with post-editing temporal and technical effort. The post-edited subtitles serve as reference and the automatic ones as hypothesis. We perform the analysis at the level of subtitles (for the metrics possible to compute at subtitle-level) and at the level of individual tasks.

text	original_text	start	orig_start	end	orig_end	time_activity
aber einander in den letzten 10 Jahren höchstens E-Mails und Statusberichte	wir senden einander E-Mails und Statusberichte	375.92	375.93	379.64	378.32	144649
	in den letzten zehn Jahren		378.41		380.05	
geschickt hatten.		379.72		380.85		12957

Table 1: Example of a process log. The first subtitle is an automatic subtitle which was edited, the second is deleted by the subtitler, while the third one is a new subtitle added by the subtitler. Time activity in milliseconds.

3.1. Effort measures

We implement the following measures of effort:
Post-Editing Speed (PES): a measure of productivity calculated as the average number of edited words per minute. Minutes are obtained from the time activity (TA) in the process logs, which measures the cumulative time spent by the subtitler on one subtitle. In practice, TA corresponds to the time that the subtitler remains 'active' on a subtitle. Consequently, TA includes operations related to editing the subtitle text, adjusting the timestamps, playing the video, but also the time spent on the subtitle without performing any specific operation.
Total interaction events (Tot_int): a measure of technical effort calculated as the sum of all keystrokes and mouse clicks performed inside the Matesub environment per task. Since the videos of each task have slightly different durations, Tot_int is normalised by video length.
Mouse clicks/interaction events (Mouse/Int): a measure of technical effort for editing the formal aspects of the subtitles, calculated as the percentage of mouse clicks over total interaction events. Mouse clicks correspond mainly to the spotting and segmentation operations, such as adding, deleting subtitles and adjusting their spotting, as opposed to editing the text.

3.2. Automatic metrics

We compute the correlation of PE speed and technical effort with the following metrics:

- **Subtitle Edit Rate (SubER)** (Wilken et al., 2022): a metric based on edit distance which considers text edits together with edits in spotting (timestamps) and segmentation. We compute SubER and its cased variant (SubER-cased).
- **Timed-BLEU (T-BLEU)** (Cherry et al., 2021): BLEU calculated over time-aligned hypothesis-to-reference. It uses linear interpolation of timestamps to penalise mistimed words.
- **AS-BLEU** (Matusov et al., 2005): BLEU computed over hypothesis-reference alignment by minimising Levenshtein distance.
- **Sigma** (Karakanta et al., 2022): a metric for evaluating the segmentation of the translated text into subtitles, irrespective of translation quality.
- **Traditional MT metrics**: Human-targeted Translation Edit Rate (HTER) (Snover et al., 2006), BLEU (Papineni et al., 2002), charF (Popović, 2015)
- **Neural metrics**: COMET (Rei et al., 2020), BERTScore (Zhang et al., 2020) and BLEURT (Sellam et al., 2020).
- **Synchronisation effort (dSpot)**: To measure auto-spotting accuracy, we compute *dSpot* as the total difference between the original timestamps of an automatically generated subtitle and the final timestamps of the edited subtitle (in seconds). *dSpot* is the sum of dStart and dEnd, where dStart is the absolute difference between the original start time of subtitle and the edited start, and dEnd is the absolute difference between the original end time of subtitle and the edited end.

3.3. Subtitle vs task processing

Since to date no subtitling tool logs keystrokes, aligning the collected keystrokes per task to individual subtitles is not possible without massive annotation effort. For this reason, at subtitle-level we only report correlations with PES based on the time activity from the process logs. The metrics that can be computed at the subtitle level are SubER, dSpot, traditional and neural MT metrics. In the process logs, original and final subtitles are already aligned, so T-BLEU and AS-BLEU overlap with sentence BLEU. We use the sentence-level versions of BLEU, chrF and TER from sacreBLEU (Post, 2018). Neural metrics are computed using MATEO (Vanroy et al., 2023).[4] Computing COMET requires the source but in subtitling the source text is audio, thus in real-life applications the written transcript may not always be available. To overcome this barrier, we back-translate the post-edited subtitles into English using Google Translate and input them as source text. Details on motivating this choice are given in Appendix B. For SubER, subtitle pairs are reconstructed in .srt format from the process logs. For added (new) subtitles, SubER is set to maximum effort (SubER=100).

[4]All signatures can be found in Appendix A.

Subtitle-level correlation may be useful for practical applications but has some drawbacks. Automatic metrics often fail for very short segments or are overly affected by reference length (see TER). Moreover, deleted subtitles have to be dropped because no time activity is recorded for them in the process logs. For these reasons, we additionally report task-level correlation, where scores are computed using the entire files per task and not per subtitle. Traditional MT metrics, neural metrics and Sigma are computed on the manually aligned text files containing one sentence per line with subtitle breaks. SubER, T-BLEU and AS-BLEU are computed on pairs of .srt files using the SubER toolkit.[5] We report two variants of T-BLEU and AS-BLEU, one considering segmentation (seg), where subtitle breaks count as additional tokens, and one ignoring segmentation.

For subtitle-level correlation, each data point represents a subtitle, while for task-level correlation it represents a task. Since the data does not satisfy linearity assumptions, we use Spearman's rank correlation coefficient.We observe some variation between subtitlers, but the shape of the distribution for PES, Tot_int and Mouse/Int is similar for all post-editors. Therefore, in addition to reporting correlations per subtitler at the subtitle level, we report correlation when concatenating the data from all three subtitlers. For talk-level correlation, the sparsity of data does not allow for obtaining meaningful correlations per subtitler, thus we only report correlations when concatenating the data.

4. Results

4.1. Productivity

Post-editing effort measurements for each of the three subtitlers for the entire data set are shown in Table 2. Italian subtitles required less edits than German, as shown by a lower HTER and total number of interactions, resulting to a higher PES for the Italian subtitlers. We observe individual differences between the two Italian subtitlers. It2 made more edits, both in terms of text (HTER) and spotting (dSpot), at a lower speed. The use of mouse also varies among subtitlers, with it1, who made the least edits having the largest ratio of mouse clicks. These individual differences are common in subtitling, where the editing process requires a complex combination of operations, and allow us to additionally investigate how metric correlations relate to different types and degrees of editing.

[5]https://github.com/apptek/SubER

	HTER	PES (w/min)	dSpot (secs)	Tot_Int	M/Int (%)
de	54.1	14.8	0.48	59,732	20.7
it1	37.8	22.1	0.68	48,336	27.8
it2	47.6	19.4	0.72	57,546	18.9

Table 2: HTER, Post-Editing Speed (PES), synchronisation effort (dSpot), total number of interaction events (Tot_Int) and ratio of mouse clicks over total interactions (M/Int) per subtitler.

4.2. Subtitle-level correlations

Correlations with PES at the level of subtitles per subtitler and aggregated for all subtitlers are shown in Table 3. In terms of individual correlations, the lowest correlations are observed for it1, who made the least number of edits at the highest PES. The correlation with dSpot, even though weak, is the highest for it1, who had the largest ratio of mouse/total interactions (27.8%). The highest correlations are noted for it2, who recorded a larger number of edits than it1 and the lowest ratio of mouse clicks. Despite having the lowest PES, the correlations for the German subtitler are between the two Italian subtitlers. Aggregating the data from all subtitlers leads to small changes in the correlation values but the rankings remain the same to a large extent.

We observe moderate correlations for most metrics. **Edit distance metrics correlate well with PES**, with HTER having the highest correlation (-0.585), followed by the cased SubER variant (-0.573) and the non-cased one (-0.562). Contrary to previous observations (Wilken et al., 2022), the higher correlations for HTER and SubER-cased suggest that considering case and punctuation increases the correlations with effort. BLEU comes next (0.541), possibly benefiting from some surface overlap between automatic and post-edited subtitles, while chrF, despite being based on characters, has the lowest correlation among all metrics. The neural metrics also obtain correlations above 0.5, showing that, despite the limited context in evaluating subtitles as short text fragments, they come close to edit-based metrics in their correlation with productivity. The moderate correlations show that **PES at such a small granularity is hard to predict**, as there is a lot of variation due to video and source text properties, among other factors.

4.3. Talk-level correlations

Table 4 shows the correlations when metrics are computed per talk. Here, all correlations increase compared to the subtitle level. This could be a result of sentence-level scoring which helps obtain more informative scores. Contrary to subtitle-level

	de	it1	it2	all
SubER	-0.592	-0.538	-0.609	-0.562
SubERcased	-0.588	-0.574	-0.619	-0.573
HTER	-0.598	-0.586	-0.622	-0.585
BLEU	0.547	0.498	0.586	0.541
chrF	0.507	0.450	0.550	0.498
BERTScore	0.529	0.502	0.583	0.536
BLEURT	0.489	0.482	0.566	0.501
COMET	0.516	0.472	0.535	0.512
dSpot	-0.334	-0.349	-0.342	-0.279

Table 3: Spearman's ρ correlation with PES per subtitler and after aggregating all data (all). All correlations are statistically significant with p<0.001.

correlations, **neural metrics obtain the highest correlations with PES**, with BLEURT and COMET scoring 0.676 and 0.641 respectively, followed by BLEU (0.659) and HTER (-0.617). The higher correlations for non-subtitle metrics may also be due to the manual sentence alignment, while subtitle metrics could be penalised by the automatic alignment.

In terms of technical effort (Tot_int), the picture is different. **Edit distance metrics have the highest correlations with total interactions**, showing that edit distance is a good proxy for technical effort. SubER here shows its value, with a correlation of 0.8. Among the BLEU variants, the metrics that consider segmentation (AS-BLEU-seg and t-BLEU-seg) have higher correlations than their segmentation-unaware counterparts. **Neural metrics obtain low correlations with technical effort compared to string-based metrics**, with only COMET having a ρ lower than -0.6. The discrepancy between PES and Tot_int is reflected in the weak correlation between PES and Tot_int (-0.33), leading to the conclusion that temporal and technical effort are different aspects of the PE process. In subtitling, many interactions are fast (mouse clicks, repeated keyboard clicks to navigate the subtitle text), which to not add to the temporal effort. On the contrary, accessing the source text requires playing the video, a factor which adds to the temporal effort but not to the technical.

When it comes to the ratio of mouse by total interaction events, **effort in editing the technical aspects of subtitling is better captured by metrics considering timings and/or segmentation**. The segmentation-aware variants correlate higher, with the time-alignment BLEU-seg having a ρ of 0.7. Casing has an effect, as shown by the higher correlation of the cased variant of SubER (-0.68 vs. -0.628). Non-subtitle string metrics (HTER, BLEU, chrF) still have moderate to high correlations, but **neural metrics only faintly capture effort of editing the technical aspects**.

Sigma does not obtain statistically significant correlations with any effort measures. This is expected, since Sigma assesses only one aspect of subtitle quality (segmentation) and does not account for the complexity of the subtitling process. dSpot shows a moderate correlation only with PES (-0.498), but we found that it also correlates with the total number of mouse operations per task (0.4501, p=0.0059) and the total time activity per task (0.5499, p=0.0005). This shows that **dSpot could be a product-based estimator for the total temporal and technical effort in editing the auto-spotting**.

Metric	PES	Tot_int	M/Int
SubER	-0.577***	0.804***	-0.628***
SubER-cased	-0.588***	0.788***	-0.680***
HTER	-0.617***	0.810 ***	-0.602***
AS-BLEU-seg	0.637***	-0.773***	0.620***
AS-BLEU	0.634***	-0.746***	0.570***
t-BLEU-seg	0.524***	-0.753***	0.700***
t-BLEU	0.601***	-0.732***	0.607***
BLEU	0.659***	-0.761***	0.608***
chrF	0.513***	-0.759***	0.615***
BERTScore	0.540***	-0.477**	0.494**
BLEURT	0.676***	-0.555***	0.378*
COMET	0.641***	-0.633***	0.423**
Sigma	0.227	-0.275	0.285
dSpot	-0.498**	-0.058	0.075

Table 4: Spearman's ρ correlation with Post-Editing Speed (PES), total number of interaction events (Tot_Int) and ratio of mouse clicks over total interactions (M/Int). Statistical significance: * p<0.05, ** p<0.01, *** p<0.001.

5. Conclusion

We presented a new corpus containing process-, product- and participant-based data of PE in automatic subtitling. The obtained correlations of automatic MT quality metrics with technical and temporal PE effort showed that edit distance metrics correlate extremely well with the total technical effort in editing automatic subtitles when considering an entire task (video). Neural metrics, when computed at the level of sentences, correlate well with PE speed, despite not considering all subtitling aspects (e.g. spotting edits). However, automatic metrics only moderately capture productivity and effort at the subtitle level, as shown by the lower subtitle-level correlations. We conclude that evaluation benefits from extended subtitle context and from considering all aspects of subtitling, including translation, spotting and segmentation. Due to the limited scope of this study, further investigations with more languages, subtitlers and domains will grant us a better understanding into the subtitle PE process, individual subtitler differences and the evaluation of automatic subtitling.

6. Acknowledgements

This project was partially funded by the EAMT programme “2021 Sponsorship of Activities - Students’ edition”. We kindly thank the subtitlers Giulia Donati, Paolo Pilati and Anastassia Friedrich for their participation in the PE task.

7. Ethical considerations

All subtitlers were properly remunerated based on their usual rates.

8. Bibliographical References

Mary Carroll and Jan Ivarsson. 1998. *Code of Good Subtitling Practice*. Simrishamn: TransEdit.

Colin Cherry, Naveen Arivazhagan, Dirk Padfield, and Krikunm Maxim. 2021. Subtitle translation as markup translation. In *Proceedings of Interspeech 2021*, pages 2237–2241.

Sheila C. M. de Sousa, Wilker Aziz, and Lucia Specia. 2011. Assessing the Post-Editing Effort for Automatic and Semi-Automatic Translations of DVD Subtitles. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 97–103, Hissar, Bulgaria. Association for Computational Linguistics.

Thierry Etchegoyhen, Lindsay Bywood, Mark Fishel, Panayota Georgakopoulou, Jie Jiang, Gerard van Loenhout, Arantza del Pozo, Mirjam Sepesy Maučec, Anja Turner, and Martin Volk. 2014. Machine translation for subtitling: A large-scale evaluation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*, pages 46–53, Reykjavik, Iceland.

Jie Huang and Jianhua Wang. 2023. Post-editing machine translated subtitles: examining the effects of non-verbal input on student translators’ effort. *Perspectives, Studies in Translatology*, 31:620–640.

Alina Karakanta, François Buet, Mauro Cettolo, and François Yvon. 2022. Evaluating subtitle segmentation for end-to-end generation systems. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3069–3078, Marseille, France. European Language Resources Association.

Maarit Koponen, Umut Sulubacak, Kaisa Vitikainen, and Jörg Tiedemann. 2020a. MT for subtitling: Investigating professional translators’ user experience and feedback. In *Proceedings of 1st Workshop on Post-Editing in Modern-Day Translation*, pages 79–92, Virtual. Association for Machine Translation in the Americas.

Maarit Koponen, Umut Sulubacak, Kaisa Vitikainen, and Jörg Tiedemann. 2020b. MT for subtitling: User evaluation of post-editing productivity. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 115–124, Lisboa, Portugal. European Association for Machine Translation.

Mariëlle Leijten and Luuk Van Waes. 2013. Keystroke logging in writing research: Using inputlog to analyze writing processes. *Written Communication*, 30:358–392.

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*.

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.

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.

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. 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.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Bram Vanroy, Arda Tezcan, and Lieve Macken. 2023. MATEO: MAchine Translation Evaluation Online. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 499–500. European Association for Machine Translation (EAMT).

Martin Volk, Rico Sennrich, Christian Hardmeier, and Frida Tidström. 2010. Machine translation of TV subtitles for large scale production. In *Second Joint EM+/CNGL Workshop*, pages 53–62.

Patrick Wilken, Panayota Georgakopoulou, and Evgeny Matusov. 2022. SubER - a metric for automatic evaluation of subtitle quality. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 1–10, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

9. Language Resource References

Alina Karakanta and Matteo Negri and Marco Turchi. 2020. *MuST-Cinema: a Speech-to-Subtitles corpus*. European Language Resources Association.

A. Data processing

We collected 3605 subtitles in total, out of which 3,019 subtitles were edited (84%), 400 deleted (11%), and 186 (5%) new. From the analysis of the screen recordings we identified that time activity was unreasonably long for some subtitles, usually subtitles at the beginning and end of the talk, or subtitles before and after a break. This suggests that the subtitlers left the project window open before starting or after finishing the task, without editing the subtitles. We thus removed these outliers, that is subtitles for which time activity was >400,000 milliseconds and time activity normalised by the number of words >20,000 milliseconds (53 subtitles), resulting in 3,552 subtitles.

The signatures for the automatic metrics are:

BLEU+case.mixed+numrefs.1+smooth.exp+tok.13a+v.1.5.0

chrF2+numchars.6+space.false+v.1.5.0

TER+tok.tercom-nonorm-punct-noasian-uncased+v.1.5.0

bertscore: nrefs:1|bs:1000|seed:12345|l:other|v:0.3.12|mateo:1.1.3

bleurt: nrefs:1|bs:1000|seed:12345|c:BLEURT-20-D12|v:commit cebe7e6|mateo:1.1.3

comet: nrefs:1|bs:1000|seed:12345|c:Unbabel/wmt22-comet-da|v:2.0.1|mateo:1.1.3

B. COMET source

One challenge in the translation of spoken texts is that the source text may not be available in written form. This creates problems in evaluation using COMET, since it requires the source. To overcome this problem, we tested two approaches: a) input the reference as source, or b) back-translate the reference and input the back-translated text as source.

To test these approaches, we compared COMET values when inputting as source either the reference or the back-translated subtitles against the true source text. We selected 80 German subtitles (MT and PE). These were back-translated into English using Google Translate. To obtain the true source text, we manually aligned the official English transcription with the target subtitles. Table 5 shows the COMET values and the mean absolute error between the COMET values when inputting the true source against those when inputting the back-translated subtitles into English or the German reference subtitles. Since the MAE is lower for inputting the back-translated subtitles, we adopted this approach in the computations of COMET.

	COMET	MAE
SRC	76.7	-
BT	75.9	0.69
REF	76.2	1.17

Table 5: COMET values for the 80 selected subtitles when inputting the true source (SRC), the back-translated reference subtitles (BT) and the German reference (REF) and the Mean Absolute Error (MAE) between SRC and BT/REF.