A Case Study on Contextual Machine Translation in a Professional Scenario of Subtitling

♣♦Sebastian Vincent, ♦Charlotte Prescott ♦Chris Bayliss,
♦Chris Oakley, ♣Carolina Scarton

♣Department of Computer Science, University of Sheffield, UK
♦ZOO Digital Group PLC, UK

Abstract

Incorporating extra-textual context such as film metadata into the machine translation (MT) pipeline can enhance translation quality, as indicated by automatic evaluation in recent work. However, the positive impact of such systems in industry remains unproven. We report on an industrial case study carried out to investigate the benefit of MT in a professional scenario of translating TV subtitles with a focus on how leveraging extra-textual context impacts post-editing. We found that post-editors marked significantly fewer context-related errors when correcting the outputs of MTCUE, the context-aware model, as opposed to non-contextual models. We also present the results of a survey of the employed post-editors, which highlights contextual inadequacy as a significant gap consistently observed in MT. Our findings strengthen the motivation for further work within fully contextual MT.

1 Introduction

As an innovation-driven company offering dubbing and subtitling services, ZOO Digital is dedicated to exploring assistive technologies to streamline our workflows. Machine translation in particular is a promising tool for improving the efficiency of the (currently fully manual) translation of the transcribed video content during interlingual subtitling. Our domain is characterised by specific challenges, both linguistic (preservation of style and function in dialogue) and practical (keeping within subtitle constraints, such as visual properties and considerations for the viewers' reading speed). We report on a case study where translation from scratch was replaced with post-editing machine translations of the source text. While such a formulation is far from new – MT has been consistently demonstrated to help reduce effort in the subtitling domain (C. M. de Sousa et al., 2011; Huang and Wang, 2023) – previous studies have relied on off-the-shelf general-purpose neural machine translation (NMT) engines like Google Translate[1]. Our work investigates two additional systems: BASE-NMT, a specialised engine trained on our data, as well its contextual version based on the MTCUE architecture (Vincent et al., 2023), whose training involves observing a vast range of metadata and document-level information.

The study was carried out with the assistance of translation and post-editing professionals. Hereinafter we refer as *post-editors (PEs)* to those who were tasked with post-editing work, and as *translators (HTs)* to those who were tasked with translation from scratch (FST). The campaign took place in a full-context multi-modal environment where the professionals had access to the video material and were able to directly jump to the segment corresponding to the utterance they were reviewing, as well as see the preceding and succeeding segments. A total of eight PEs were employed, four for English-to-German (EN-DE) and four for English-to-French (EN-FR) translation, and four HTs, two per language pair. We measured the effort it took to post-edit or translate the TV series content and the number of specific translation errors observed by the PEs. Our findings highlight

[1] https://translate.google.com/

the necessity of tailoring MT engines to the target domain and motivate further work within leveraging contextual systems in dialogue translation.

2 Related Work

Over the last few years, subtitle translation has been given a volume of attention: C. M. de Sousa et al. (2011), Koponen et al. (2020) and Huang and Wang (2023) observe that post-editing the outputs of an NMT system is a promising alternative to translation *ex novo*, reducing the temporal, technical and cognitive effort of both novice and professional translators and subtitlers. A survey among professional subtitlers detailed by Karakanta et al. (2022), finds that professionals have a positive outlook on incorporating automatic components (such as MT) into their workflow, as they offer starting templates, reduce effort and can provide useful suggestions. However, some challenges in the automatic translation of subtitles remain unsolved (Gupta et al., 2019; Karakanta et al., 2022), including the adherence to subtitle block limitations, which often necessitates shorter and paraphrased translations; lexical consistency, which involves translating the same terms across the text, as well as using vocabulary that maintains the cohesion and coherence of the text, aligns with the surrounding video or textual content, and conforms to standard language or industry conventions; lexical errors such as the translation of idioms and figurative language, and context-related inconsistencies. Context-related errors in particular have been pointed out as the culprit in many works in MT that leveraged the OpenSubtitles corpus (Lison et al., 2018), a dataset of user-submitted subtitles and their translations. Leveraging document-level information (Tiedemann and Scherrer, 2017; Bawden et al., 2018), speaker's and interlocutor's gender identity (Vincent et al., 2022) and explicit extra-textual information (Vincent et al., 2023) has been found particularly useful in addressing this challenge. Context is also useful during the manual post-editing procedure: Huang and Wang (2023) show that such a setup decreases the cognitive load of student translators compared to a text-only scenario, suggesting as an explanation the dual coding theory, according to which the interactions between the verbal and non-verbal information enhances the translators' understanding of the material.

This work employs MTCUE (Vincent et al., 2023), a multi-encoder Transformer designed for contextual NMT capable of leveraging contextual signals such as film metadata and document-level information to improve translation quality, as well as enabling better control of phenomena such as speaker's gender and formality register. The mechanism for delivering context in the model involves converting the context fields into equal-sized vectors via sentence embedding. The resulting vector sequence is inputted into a distinct Transformer encoder. Additionally, we employ the context specificity evaluation method outlined in Vincent et al. (2024), which relies on the pointwise mutual information (PMI). In this method, PMI quantifies the degree of co-occurrence between tokens in a translation hypothesis and the respective context.

3 Experimental Setup

The primary objective of our case study was to investigate whether post-editing MT is a cost-effective alternative to FST in our workflow, and to what extent domain-adapted training data and the utilisation of context have an impact in this area. Guided by the availability of resources, we operated in two language pairs: EN-DE and EN-FR and considered four versions of the text in each, including MT outputs from three systems:

1. GOOGLE[2], a general-purpose NMT engine used in previous work.
2. BASE-NMT, a non-contextual Transformer-based translation model parameter-matched to MTCUE and trained on the same data (except context).
3. MTCUE system (Vincent et al., 2023), a multi-encoder Transformer.

We also operated on the human translations of the test set (REF) approved for production.[3]. For both MTCUE and BASE-NMT, we trained the models after Vincent et al. (2024), §4.1, in the OVERLAP setting which mimics a scenario with access to prior episodes of a tested series for training (a sample is presented in Appendix F of that work). We operated on sentence-level translations, with MTCUE using the context for each sentence in its dedicated space.

[2] https://translate.google.com/

[3] This baseline is omitted during automatic evaluation (in fact, it is used as the reference text to calculate the automatic metrics), but is used as a baseline in the human evaluation, where the professionals are asked to post-edit this already sufficiently good text.

3.1 Automatic evaluation

We conducted a pre-emptive automatic evaluation to confirm the feasibility of the human evaluation study. We used BLEU (Papineni et al., 2002) and COMET (Rei et al., 2020) as translation quality metrics. Additionally, to measure context specificity, we measured the PMI between contextual and non-contextual translations (Vincent et al., 2024). We compared the outputs of the machine translation systems (BASE-NMT, GOOGLE, MTCUE) against the reference (REF).

3.2 Post-Editing Setup and Metrics

The human evaluation aspect of the study is interpreted as the effort required to post-edit the translations to a production standard, and captured in the **number of errors, keystrokes** and **total edit time**. The task was performed by professional HTs and PEs using ZOOSUBS, an in-house software application belonging to ZOO Digital, built to facilitate manual translation of video material (Figure 1). The software's interface displays the video material along with timed subtitles in the original language. The *target stream*, i.e. the set of text boxes provided to the right of the source stream, is where the HTs input their translations to the desired language. It can optionally be pre-populated with "draft" translations – a setting we opted for in this study – allowing post-editors to edit, divide or combine the segments as they see fit.

To make amendments to a segment, the PE needs to click on its box. From that point, the system tracks the time spent editing the box and the number of keystrokes made. These metrics are recorded for each box separately and taken into account only if the post-edited text differs from the original. After applying modifications, an **Issues for event** window appears for the user to specify the purpose of the changes by selecting errors from a predefined list, optionally providing text commentary. We leveraged this functionality of ZOOSUBS to measure the total and average time and number of keystrokes made by HTs and PEs given some pre-existing translations. We also measured the number of selected errors. For this project, we created a bespoke taxonomy of errors (Table 1) based on translation errors reported in previous work (Freitag et al., 2021; Sharou and Specia, 2022), the original list of issues already present in the ZOOSUBS system and relevant errors from previous work (§2). Error categories from the aforementioned sources were compiled together and curated to fit the study requirements[4]

Worker setup The PEs operated on seven episodes from three TV series of varying genres: a fictional series about space exploration, a documentary exploring aspects of everyday life, and a family cooking competition show. They were unaware that some of the text they worked with was machine translated, but were told that it was for a research project and asked to relax some constraints such as adhering to the reading speed limits. In addition, we asked four HTs (two to German, two to French) to translate one episode of the cooking show from scratch in ZOOSUBS so we could compare their effort to that of post-editors. For each of the seven episodes, the PEs were asked to post-edit one out of four versions of the text, corresponding to the list outlined in §3. We included the human references (REF) to account for the fact that PEs can sometimes post-edit a translation even when the original one is valid. Our setup ensured that the same PE evaluated the output for each episode exactly once (i.e. does not see two different versions of the same text) (Table 2). When referring to individual PEs, we use the notation **PE.[L][i]**, where $\mathbf{L} \in \{\mathbf{G}$ (German), **F** (French)$\}$, and **i** denotes the PE ID $\in [1, 4]$.

Details regarding the PEs The recruited PEs and HTs were professionals within the subtitle domain and freelance employees of ZOO DIGITAL. They were informed that the undertaken work was carried out for a research project, but nevertheless, they were paid for their effort at competitive PE and HT rates, standard within the company for this type of work. Information about the PEs' and HTs' years of experience (YOE) was collected to shed more light on the findings (Table 3). They also answered a short survey about their views regarding machine translation, discussed in detail in §5.3:

1. Which one would you prefer: translating a stream from scratch or completing a quality check on (post-editing) a stream? Why?
2. What are your views on the use of machine translation in the industry?
3. In your view, are there benefits to post-editing translations over translating from scratch?

[4]We uploaded a draft taxonomy to ZOOSUBS, and the first author performed a test evaluation against a stream with 446 segments to validate the list. As a result, some errors were split into more granular categories, some were renamed and some generalised.

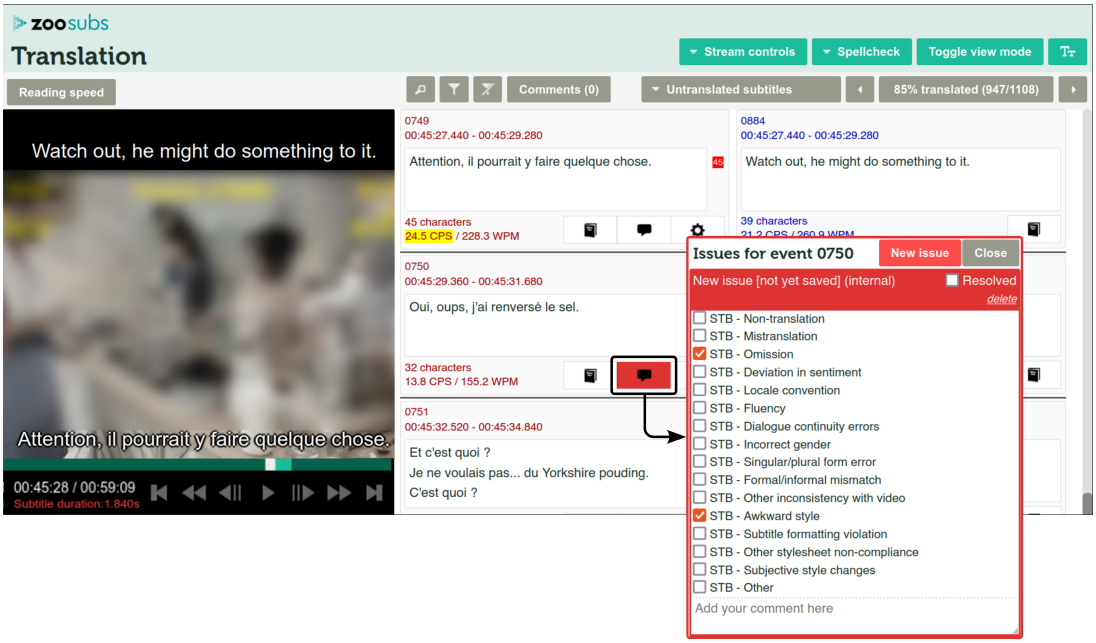

Figure 1: A compressed snapshot of ZOOSUBS.

Type	Description
Translation quality	
Catastrophic translation	Impossible to post-edit, must be translated from scratch.
Mistranslation	Incorrect. Does not preserve the meaning or function of the source.
Omission	Part of the source text was left untranslated.
Deviation in sentiment	Does not preserve the sentiment of the source (e.g. does not match the expressed excitement), or negates the sentiment (e.g. from positive to negative).
Locale convention	Violates locale convention, e.g. currency and date format.
Fluency	Contains punctuation, spelling and grammar errors.
Context	
Incorrect gender	Misgenders the speaker or the addressed person(s).
Incorrect plurality	Incorrectly refers to a single person when a group is addressed, or vice versa.
Wrong formality	Expressed in informal style or uses informal addressing when should use formal, or vice versa.
Other inconsistency with video	Contains inconsistencies with the video material not falling within any of the above.
Style	
Subtitle formatting violation	Violation of the subtitle blocking guidelines.
Other style sheet non-compliance	Does not conform to the provided style sheet.
Awkward style	The style of the translation does not reflect the style of the source sentence and/or the context.
Subjective style changes	The translation is acceptable but the editor suggests improvements in style.
Other	Error of type not found above (use text box provided).

Table 1: List of errors provided to the human evaluators during the campaign.

Series	A		B		C		
Ep. ID	A1	A2	B1	B2	C1	C2	C3
PE.1	REF	MTCUE	GOOGLE	BASE-NMT	REF	MTCUE	GOOGLE
PE.2	BASE-NMT	REF	MTCUE	GOOGLE	BASE-NMT	REF	MTCUE
PE.3	GOOGLE	BASE-NMT	REF	MTCUE	GOOGLE	BASE-NMT	REF
PE.4	MTCUE	GOOGLE	BASE-NMT	REF	MTCUE	GOOGLE	BASE-NMT
HT.1	*From Scratch*						
HT.2	*From Scratch*						

Table 2: Work assignment to PEs and HTs in the human evaluation campaign used for both language pairs.

	English-to-French				English-to-German			
	PE.F1	**PE.F2**	**PE.F3**	**PE.F4**	**PE.G1**	**PE.G2**	**PE.G3**	**PE.G4**
Translation YOE	15	8	3	20	7	18	8	17
YOE in subtitles	8	6	1.5	20	7	5	8	7
YOE in post-editing	8	6	3	10	5	5	1	3
Post-editing training?	✓	✓	✓	✓	✗	✗	✗	✗
Prefer post-editing?	✓	✓	✗	✓	✓/✗	✗	✗	✗

Table 3: Details regarding employed PEs.

All French HTs had training in post-editing, and three out of four preferred it to translating from scratch, while no German HTs had received such training in the past, and all but one strictly preferred FST. All PEs had at least one YOE in post-editing and one and a half in the subtitle domain. Although the HTs within both pairs had a similar

amount of experience in translation in general and in the subtitle domain (11.5 ± 6.5 for French vs 12.5 ± 5.0 for German), the French HTs had the advantage in terms of YOE in both subtitling (a mean difference of 2.1 YOE) and post-editing (a mean difference of 3.3 YOE).

4 Results of Automatic Evaluation

The automatic evaluation results (Figure 2) suggest that MTCUE was the best-performing system and GOOGLE the worst-performing for both language pairs. Interestingly, for EN-DE, the BLEU and COMET score differences varied in magnitude, to the point of COMET judging all three systems as on par. A possible cause was the discrepancy in hypothesis length (the reference text uses 7.04 words per segment, BASE-NMT: 7.06, MTCUE: 7.06, GOOGLE: 8.29). Since COMET's calculation involves comparing sentence embeddings of the hypothesis and the reference, including more words or phrases in the hypothesis may lead to a closer similarity match, inflating the score even if the additional tokens are redundant or even harmful to quality. BLEU does not have this problem as it is based on string matching (Papineni et al., 2002). As per the PMI scores, the professional translations (REF) consistently exhibited the highest context specificity. However, MTCUE was on par with this reference score in both cases and was consistently better than the other two systems. MTCUE therefore shows promise at addressing the context-related issues in subtitle translation.

5 Results of the Post-Editing Study

This section analyses the results of the post-editing study: the translation errors (§5.1), the post-editing effort (§5.2), and finally, the post-campaign survey responses (§5.3).

Due to the unprecedented nature of this work in the company, the professionals' contract allowed them to withdraw if they found the compensation insufficient for the requested work. At the midpoint of the campaign, two PEs (**PE.G1** and **PE.G3**) contacted the project manager to express concerns regarding the quality of the MT outputs, asserting that the task potentially required more effort than FST. To compromise, they proposed narrowing the scope of the remaining work to error identification and marking, without making the necessary corrections. This meant we would not obtain the effort metrics for the two PEs. Consequently, while the error analysis in §5.1 includes both language pairs, the effort analysis in §5.2 does not include results from **PE.G1** or **PE.G3**.

5.1 Error Analysis

An initial inspection of the results indicated that each PE marked a significantly different total number of errors (e.g. **PE.F1** marked 232 errors total while **PE.F4** marked 878). This made direct comparison of the error counts across systems unreliable as each PE also post-edited a different number of segments for each system (cf. Table 2). With seven episodes and four different versions of the text, for each PE there is a version of text they would only have seen one episode from. For example, in Table 2, **PE.1** is assigned two episodes for REF, MTCUE and GOOGLE, but only one for BASE-NMT. In this example, if **PE.1** generally marked fewer errors than others, BASE-NMT would be disproportionately rewarded.

To make the measurements comparable, we normalised them by computing a *normalisation coefficient* h for each PE and then multiplying their error counts for each category by their h. Let $\text{ERR}_{PE_i,c}$ denote the number of errors within the category c for the i-th PE. We compute the normalised count $\widehat{\text{ERR}}_{PE_i,c}$ as described by Equation 1.

$$\widehat{\text{ERR}}_{PE_i,c} = \text{ERR}_{PE_i,c} \times h_i$$
$$\text{where } h_i = \frac{\max(\text{ERR}_{PE_j,total}; j \in \{1,4\})}{\text{ERR}_{PE_i,total}} \quad (1)$$

We report the total error counts as well as the normalisation multipliers in Table 4.

English-to-German			English-to-French		
PE ID	*Error count*	h	*PE ID*	*Error count*	h
PE.G1	1526	1.76	**PE.F1**	232	14.68
PE.G2	2452	1.10	**PE.F2**	182	18.71
PE.G3	2690	1.0	**PE.F3**	3406	1.0
PE.G4	1832	1.47	**PE.F4**	878	3.88

Table 4: Error counts and values of h for each PE.

Error post-processing To facilitate post-editing in ZOOSUBS, MT outputs had to be adapted to match the subtitle format. Quality checks of translations conducted in ZOOSUBS normally require the users not just to ensure the correctness of translations but also that the subtitles comply with strict guidelines[5]. Typical MT systems, like the ones

[5]This includes adhering to reading speed and length limits, balancing the length of the top and bottom subtitle, disam-

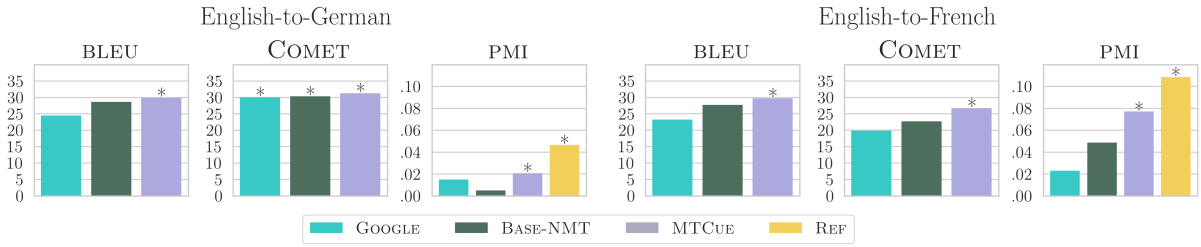

Figure 2: BLEU, COMET and PMI scores obtained by the evaluated models. Asterisks (*) over bars indicate the best result along with all statistically indistinguishable results computed either via bootstrap resampling (or t-test for PMI), $p = 0.05$.

used in this project, are not designed to create translations conforming to these stringent guidelines, and the primary goal of this study was to identify the impact of the translation errors alone. To faithfully replicate the normal work environment of the PEs, we applied a greedy reformatting tool (built into ZOOSUBS) to reformat our translations as subtitles. We made it clear that the project is centred on the correctness of translations, not the subtitle formatting. Still, to ensure that the translation and non-translation errors are kept separate, we included two environment-specific errors for the workers to select from: *Subtitle formatting violation* covering cases where the subtitle is not split to optimally adhere to segmentation guidelines; and *Other style sheet non-compliance* where a rule outlined in the style sheet from the client company was not followed, such as custom punctuation conventions.

Example 1	Target: German
Source	Can I take a look at what you're doing by any chance?
BASE-NMT (✗)	Kann ich mir zufällig ansehen, was du [BR] machst?
Post-ed.	Kann ich mir vielleicht ansehen, [BR] was Sie da machen?
Errors	*Mistranslation* *Subtitle formatting violation* *Formal/informal mismatch*

In some instances, a PE would encounter both translation and non-translation errors within the same segment, as exemplified in **Example 1**, where both translation errors (*Mistranslation* of *by any chance* and *Formal/informal mismatch* of *you're doing*) and non-translation errors (*Subtitle formatting violation* of the position of the subtitle break) are present. In such cases, we (i) disregard the non-translation error counts, and (ii) correct the effort rates (editing time and keystrokes) to account solely for translation-related errors. To precisely gauge the latter, we employed a correction method: let $\text{ERR}_{non-translation}$ and $\text{ERR}_{translation}$ be the total effort expended by a PE on a segment that had only non-translation and only translation errors marked, respectively. We calculated translation share (TS) as follows:

$$\text{TS} = \frac{\text{ERR}_{translation}}{\text{ERR}_{translation} + \text{ERR}_{non-translation}}$$

We then used it to calculate the estimated share of the effort spent on translation in segments that had both errors marked by multiplying TS by the total effort spent on a segment with both error types.[6]

Finally, since the **Other** category was used substantially, we parsed the contents of the optional description text box. The most commonly reported **Other** errors were "Grammar", "Punctuation", "Timing", "SGP" (spelling, grammar, punctuation) and "Literal translation". Such errors (69.3%) were removed from the **Other** category and pigeonholed as appropriate (e.g. "Grammar" as *Fluency*). More complex comments such as "wissen Sie should not be in the translation" were left categorised as *Other* (30.7%).

Results The calculated normalised counts of errors within each category (Table 5) suggest that MTCUE performs no worse than both non-contextual MT systems overall (row **Total**), while performing significantly better in the **Context** and **Style** categories in EN-FR, pointing to gains related to the use of context information.

The most frequently flagged errors in both language pairs were consistently *Mistranslation* and *Fluency*. *Mistranslation* was reported a similar number of times for all three machine translation

biguation of speaker turns with colours or dashes, and applying appropriate formatting, as specified by a style sheet.

[6]For example, if a PE took three seconds for translation errors and two seconds for non-translation errors on average, where they marked both types we multiplied their total effort for that segment by $\frac{3}{3+2}$.

	Error type	Normalised count GOOGLE	BASE-NMT	MTCUE	REF
English-to-German	**Translation quality**	13.12 ± 14.46	$\underline{8.70 \pm 11.67}$	$\mathbf{8.49 \pm 10.90}$	4.56 ± 5.14
	Catastrophic translation	$\underline{0.50 \pm 0.27}$	$\mathbf{0.46 \pm 0.18}$	$\underline{0.88 \pm 0.95}$	0.72 ± 0.68
	Mistranslation	$\underline{26.99 \pm 8.58}$	$\mathbf{25.69 \pm 7.67}$	$\underline{26.74 \pm 6.15}$	8.76 ± 5.51
	Omission	$\mathbf{0.26 \pm 0.15}$	2.32 ± 2.20	3.54 ± 2.79	5.38 ± 6.75
	Deviation in sentiment	$\underline{1.11 \pm 0.66}$	$\mathbf{0.83 \pm 0.30}$	$\underline{1.25 \pm 0.88}$	5.23 ± 4.40
	Locale convention	2.04 ± 0.00	$\underline{0.94 \pm 0.46}$	$\mathbf{0.61 \pm 0.30}$	0.91 ± 1.03
	Fluency	16.88 ± 15.22	$\underline{9.54 \pm 11.17}$	$\mathbf{7.10 \pm 6.52}$	4.18 ± 3.65
	Context	5.34 ± 5.68	$\underline{2.64 \pm 3.45}$	$\mathbf{2.21 \pm 2.55}$	1.18 ± 1.13
	Incorrect gender	$\underline{2.20 \pm 1.58}$	$\underline{1.69 \pm 1.90}$	$\mathbf{1.43 \pm 1.17}$	1.60 ± 1.19
	Plural/singular form error	$\underline{0.99 \pm 0.81}$	$\mathbf{0.80 \pm 0.63}$	$\underline{1.19 \pm 1.24}$	0.33 ± 0.00
	Formal/informal mismatch	11.31 ± 4.55	$\underline{5.29 \pm 4.60}$	$\mathbf{3.86 \pm 3.60}$	1.19 ± 1.31
	Style	$\underline{12.19 \pm 9.79}$	$\mathbf{8.12 \pm 6.59}$	$\underline{9.88 \pm 7.83}$	3.77 ± 3.86
	Awkward style	17.70 ± 7.76	$\mathbf{11.82 \pm 5.21}$	$\underline{13.11 \pm 7.04}$	4.70 ± 4.34
	Subjective style changes	$\underline{2.55 \pm 2.09}$	$\mathbf{1.65 \pm 1.59}$	$\underline{2.33 \pm 2.28}$	2.13 ± 2.52
	Other	$\underline{2.12 \pm 3.43}$	$\underline{3.26 \pm 4.48}$	$\mathbf{2.10 \pm 2.46}$	3.39 ± 5.88
	Total	9.58 ± 11.35	$\underline{6.44 \pm 9.05}$	$\mathbf{6.41 \pm 8.82}$	3.86 ± 4.70
English-to-French	**Translation quality**	20.01 ± 23.05	$\mathbf{9.27 \pm 9.52}$	$\underline{10.21 \pm 8.88}$	6.60 ± 5.08
	Catastrophic translation	$\underline{3.41 \pm 1.38}$	$\mathbf{2.25 \pm 2.39}$	$\underline{2.86 \pm 3.03}$	2.51 ± 3.26
	Mistranslation	38.80 ± 14.35	$\underline{22.73 \pm 8.49}$	$\mathbf{20.10 \pm 7.34}$	7.24 ± 3.61
	Omission	$\mathbf{2.40 \pm 2.40}$	$\underline{3.91 \pm 1.49}$	5.56 ± 4.09	7.48 ± 5.13
	Deviation in sentiment	$\mathbf{5.93 \pm 5.90}$	$\underline{7.82 \pm 6.09}$	11.59 ± 0.00	6.74 ± 3.03
	Locale convention	4.29 ± 2.49	0.73 ± 0.51	$\mathbf{0.21 \pm 0.00}$	0.63 ± 0.00
	Fluency	30.83 ± 31.77	$\underline{7.28 \pm 3.75}$	$\mathbf{5.92 \pm 4.18}$	7.82 ± 7.35
	Context	$\underline{5.41 \pm 3.64}$	6.09 ± 4.26	$\mathbf{3.86 \pm 3.11}$	1.29 ± 1.07
	Incorrect gender	$\mathbf{3.49 \pm 2.59}$	6.96 ± 5.57	$\underline{4.77 \pm 3.98}$	0.49 ± 0.44
	Plural/singular form error	4.50 ± 1.92	5.84 ± 4.60	$\mathbf{1.97 \pm 0.62}$	0.00 ± 0.00
	Formal/informal mismatch	$\underline{7.44 \pm 4.63}$	$\underline{5.58 \pm 3.76}$	$\mathbf{4.23 \pm 2.93}$	1.69 ± 1.10
	Style	11.05 ± 7.07	10.35 ± 3.69	$\mathbf{3.41 \pm 2.53}$	5.55 ± 3.41
	Awkward style	11.13 ± 7.46	9.55 ± 1.27	$\mathbf{2.89 \pm 2.76}$	4.10 ± 1.28
	Subjective style changes	$\underline{10.94 \pm 8.16}$	11.15 ± 5.52	$\mathbf{4.18 \pm 2.87}$	6.28 ± 4.09
	Other	37.20 ± 52.68	$\mathbf{11.19 \pm 16.44}$	$\underline{23.67 \pm 29.23}$	27.05 ± 24.68
	Total	17.02 ± 25.78	$\mathbf{8.84 \pm 9.20}$	$\underline{9.63 \pm 13.85}$	8.83 ± 12.84

Table 5: Counts of errors flagged by the PEs for each system. Excluding REF, the best result in each row is highlighted and all statistically indistinguishable results are underlined (one-tailed t-test, confidence interval of 80%, $p = 0.2$). Error rates for categories in bold (e.g. **Style**) are calculated based on all errors within the category.

systems in EN-DE and three times less frequently for post-editing REF. This gap was similar in EN-FR, though within the MT systems themselves, the GOOGLE system had a significantly higher error rate for *Mistranslation* errors (38.80 mean) than the next best system, i.e. BASE-NMT (22.73); the contextual MTCUE achieved an even lower rate of 20.10. Interestingly, MTCUE also produced outputs of higher *Fluency* than other systems, even surpassing REF for EN-FR, though insignificantly at the selected confidence interval (80%).

In both language pairs, the *Omission* error was consistently marked the fewest times in GOOGLE-generated text (see **Translation quality** → *Omission*). In both cases, REF scored significantly above the mean. This is unsurprising: translations authored by the general-purpose GOOGLE engine tend to be overly literal and faithful to the source, while in the domain of dialogue, the HT often needs to let go of individual features of the source text or opt for alternative expressions to maintain the brevity and dynamics of the source dialogue, leading to spontaneous omissions in the reference translations. To exemplify, GOOGLE consistently unnecessarily translated the English *"(...), you know,"* to *"(...), wissen Sie,"* in German, necessitating additional post-editing in our study. A similar error was typically avoided by

the other systems, due to their data-learned preference for brevity and dynamically expressive language. As a result, both systems were marked with *Omission* more times than GOOGLE. In fact, MTCUE scored even more *Omission*s than BASE-NMT, suggesting that MTCUE's omission behaviour more closely matches that of professional HTs. Other **Translation quality** errors were relatively infrequent and with insignificant differences between systems.

To capture context-related issues, we provided categories for the most frequent contextual errors: *Incorrect gender*, *Plural/singular form* and *Formal/informal mismatch*. Since the perception of speaking style in dialogue is subjective and difficult to gauge, we did not provide explicit ways for the PEs to mark speaker style errors to avoid biasing them towards thinking in terms of what is a characteristic way of expression for the given speaker. Instead, we provided loose categories for **Style**, with the intention of collecting measurements of how often the PEs feel the need to alter the style of the translations. Since all of the post-edited content is dialogue, the style of the translation can be directly associated with the style of the speaker's expression. Our findings regarding some **Context** categories (*Incorrect gender*, *Formal/informal mismatch*) are consistent between the two language pairs, and MTCUE was found to be superior in most categories in both cases, with the overall score for the **Context** category being significant at 80% confidence for EN-FR. The *Plural/singular* form error required few corrections in EN-DE (where BASE-NMT was found superior to MTCUE) and more in EN-FR (where MTCUE was found superior).

The findings from the **Style** category also work in favour of contextual MT, where it was found comparable to non-contextual systems for the EN-DE pair and significantly better than them for the EN-FR pair, requiring the fewest style-based adjustments, even fewer than REF. Within the EN-DE pair, *Subjective style changes* were flagged only up to $4-5$ times per 100 segments for any system, and a consistent number of times between systems, and *Awkward style* was flagged the fewest times for REF (4.68 on average), much less frequently than for the other systems, among which GOOGLE required the most edits and BASE-NMT the fewest.

Overall, our error count analysis suggests that within the EN-FR pair, MTCUE has significantly reduced the number of errors marked for contextual and stylistic reasons compared to non-contextual systems, while not degrading overall translation quality. The findings within the EN-DE pair are too variable to yield definitive conclusions but entail no degradation of quality leading from the inclusion of context, a significant improvement for contextual phenomena compared to GOOGLE, and highlight that MTCUE makes the fewest contextual errors overall.

5.2 Analysis of Effort and Quality

This section delves into the analysis of per-PE effort spent post-editing or translating the outputs of each system. Based on the observation that some measurements of editing time and keystrokes were out of the distribution, we normalised these by first computing the 97.5th percentile for the given language pair and task (translation or post-editing) and set all per-segment measurements to be capped at that percentile. Our obtained percentiles were: 37 seconds and 69 keystrokes for translation, and 45 seconds and 54 keystrokes for post-editing.

Effort per PE As per Figure 3, the results for the EN-DE pair suggest that each PE contributed a similar effort. Interestingly, the error rate and effort measures of these PEs are closer in magnitude to the outlier **PE.F3** within the EN-FR pair. Putting PEs from both pairs together we find an interesting correlation: those PEs who expressed a preference for post-editing marked significantly fewer errors overall. We suspect that professionals who expressed a preference for translation opted for spending any effort necessary to match the quality of the resulting text to what they would have produced from scratch, while the post-editing enthusiasts contributed fixed effort, possibly characteristic of their usual post-editing assignments.

The error rate for this pair points to GOOGLE as the system consistently requiring the most edits, and REF the least, though only **PE.G4** made drastically fewer edits to this already production-ready text. Between BASE-NMT and MTCUE, **PE.G2** and **PE.G3** found MTCUE to be less erroneous (and **PE.G3** found it to be on par with REF), while **PE.G1** and **PE.G4** identified fewer errors in BASE-NMT.

According to **PE.G2**, the quality of translations from GOOGLE and BASE-NMT is comparable, requiring the most complex and laborious edits. MTCUE's hypotheses required less work

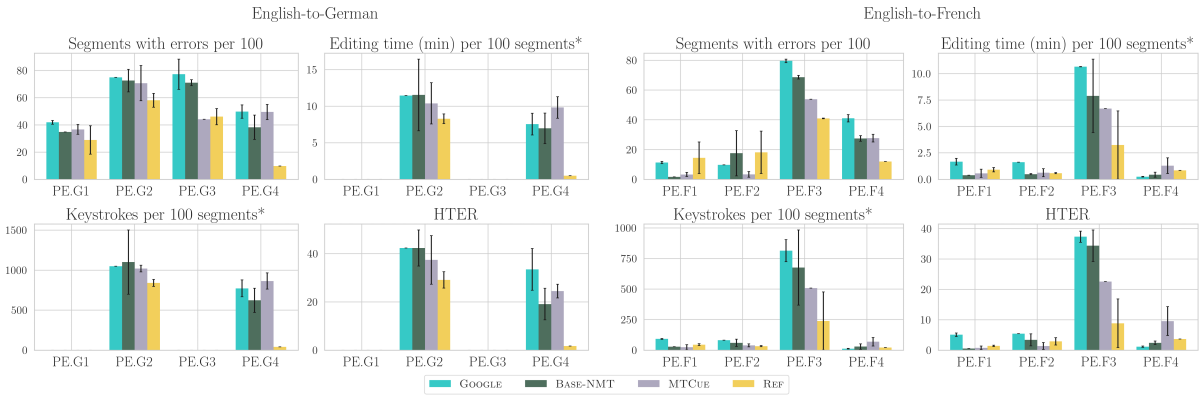

Figure 3: Effort for each PE within both language pairs.

from this PE, and REF text still less. Results obtained from **PE.G4**'s edits are different, revealing next to no edits to the REF text, (which could be interpreted as them being the least subjective of the PEs, only making edits when they are necessary). This PE found MTCUE to require more edits than BASE-NMT and on par with GOOGLE. Interestingly, even though editing MTCUE's outputs took more time and keystrokes, GOOGLE's outputs yielded a HTER value about 10 points higher than MTCUE. Since GOOGLE is the more literal MT system, and MTCUE produces more dialogue-like responses, these findings suggest that, other things being equal, a literal and overly long translation of dialogue may take less effort to post-edit than an incorrect platonic (dialogue-like) response, even if more profound edits are required.

Approach to REF Since the PEs were told about the research nature of the project, they might have approached this project with less vigilance than if the work was undertaken for actual clients. On the flip side, some may have eventually realised they were dealing with some MT outputs – they were not told this explicitly – and became more scrutinous as a result, expecting to make many more corrections than in a typical post-editing task. This would perhaps explain why some PEs took to post-editing REF at rates sometimes matching the outputs of the MT systems, with three of them doing so at a rate of over 40 errors per 100 segments.

Comparison with translation effort In Figure 4 we compare the unnormalised post-editing effort (exclusive of REF) to the FST effort for one episode of the cooking show. For both language pairs, FST required 4 to 6 times the effort of post-editing, by both measures.

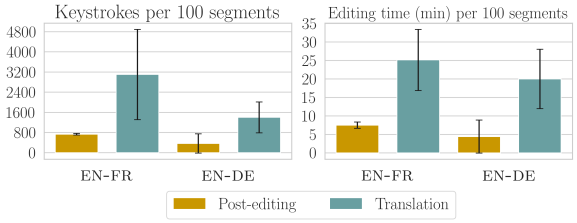

Figure 4: Effort comparison of FST and post-editing MT.

5.3 Analysis of the professionals' views on post-editing and MT

Finally, we present the PEs' responses to a survey regarding views on post-editing and machine translation. Most of the German PEs expressed a preference for FST over post-editing, with three voicing frustration with MT's stiffness and literal nature, omitting aspects of the original text such as slang, gender agreement, references to the video and people's speaking styles. They view translation as a more creative process which can yield idiomatic and fluent translations. They also noted that post-editing currently demands more effort than translating from scratch at times, yet it is compensated at a lower rate than translation. To one PE, post-editing felt like damage control.

Conversely, three out of four French PEs expressed a preference for post-editing, justifying the choice with their specialisation. The fourth PE was dissatisfied with the amount of subtitle formatting errors within our project, commenting that FST would have focused more on content.

PEs in both languages agreed that MT can be a helpful tool, and praised the recent developments, but still concurred that the substantial gap in quality persists, and renders MT insufficiently competent to replace FST. However, they were optimistic about future developments within MT. The

majority of PEs recognized the advantages of post-editing, such as the reduction of temporal effort in some cases and the potential to improve consistency in translating terminology, and enabling greater attention to detail. However, presently these benefits can fail to materialise in practice, emphasising the importance of further work on implementation quality of post-editing workflows.

5.4 Examples of challenges

We present two examples of corrections made in the post-editing process to reflect what kind of corrections required attention as well as what mistakes need to be improved upon in the future.

Example 2	Target: German
Source	No way, no way.
Video context	*The victorious family is in disbelief about their triumph.*
MTCUE (✗)	Auf keinen Fall. (‘*Under no circumstance.*’)
Post-ed.	Unmöglich. (‘*Unbelievable.*’)
Error	**Other**: *inconsistency with video*

Example 2 presents a scenario where MTCUE incorrectly interprets the exclamation *No way* as *Under no circumstance*, which fails to account for the sense of disbelief and amazement that the victorious family is experiencing. Such an interpretation relies strongly on the visual context, of which effective incorporation into the machine translation process in a multi-modal framework is an area for future work.

Example 3	Target: German
Video context	*Two cooks and a chopping board.*
Source N	Get that Welly on that board.
Reference N	Leg das Welly auf das Brett.
MTCUE (✗)	Stell die Welly auf das Brett.
Post-ed.	Legt das Wellington auf das Brett.
Error	*Awkward style*
Source N+1	She’s on.
Reference N+1	Es ist drauf.
MTCUE (✗)	Sie ist dran.
Post-ed.	Ist drauf.
Error	**Other**: *inconsistency with video*

Example 3 presents a two-error scenario. Firstly, MTCUE uses the incorrect German preposition *an/dran* to translate the English *on*, instead of the correct *auf/drauf* (*on that board* = *auf das Brett*). The more interesting error comes from mistranslating *She* as *Sie*. The pronoun is a reference to pork Wellington, abbreviated to *Welly* by the speaker, and incorrectly assigned the feminine article *sie*, instead of the neuter *das*. The speaker personifying the pork in Source N+1 (referring to it as *She*) complicates things, and so even a document-level system could have trouble interpreting what *Welly* actually is. The correct interpretation is crucial to selecting the right verb *legen* over *stellen* which should be used to translate *get* when referring to meat. Though it was marked with an *inconsistency with video* error, it is challenging to outline the minimal set of context information sufficient for the correct treatment of this example. The context of cooking, the light-hearted, casual character of the show and the manner of British speech, as well as what meal is being made and what the cooks are doing at the moment, all could aid this process. An important challenge for future contextual systems is going to be to discern which type of information is necessary and when.

6 Conclusions and Future Work

We have presented a case study on post-editing MT of subtitles for TV series in a multi-modal scenario, with a focus on contextual MT. We found that the MT models custom-trained on dialogue required less post-editing effort than the one-size-fits-all Google Translate, potentially due to the overbearing literalness and stiffness of the latter system’s outputs. We also found that some post-editors amended production-approved human translations at high rates, with hypervigilance about dealing with MT as a possible cause. Our results did not determine a significant difference in post-editing effort between MTCUE and BASE-NMT. However, the inclusion of context in MTCUE yielded fewer errors in the **Style**, **Context** and *Fluency* categories, motivating our future exploration of context-inclusive models. We further found that post-editing any MT output required four to six times less technical and temporal effort compared to FST, making it a promising cost-effective venture. However, cognitive effort should be measured in future studies, given the exit survey sentiment that post-editing was sometimes harder and less interesting than FST. Our future experiments will employ larger cohorts of PEs and split them into groups who post-edit non-contextual and contextual inputs exclusively, so that clearer feedback can be collected, as well as to minimise the variance in effort.

7 Acknowledgements

This work was completed as part of Sebastian Vincent's PhD, which was partially funded by ZOO Digital. Sebastian was also supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation (grant number EP/S023062/1).

References

[Bawden et al.2018] Bawden, Rachel, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, 1:1304–1313.

[C. M. de Sousa et al.2011] C. M. de Sousa, Sheila, 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, September. Association for Computational Linguistics.

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

[Gupta et al.2019] Gupta, Prabhakar, Mayank Sharma, Kartik Pitale, and Keshav Kumar. 2019. Problems with automating translation of movie/tv show subtitles. CoRR, abs/1909.05362.

[Huang and Wang2023] Huang, Jie and Jianhua Wang. 2023. Post-editing machine translated subtitles: examining the effects of non-verbal input on student translators' effort. Perspectives, 31(4):620–640.

[Karakanta et al.2022] Karakanta, Alina, Luisa Bentivogli, Mauro Cettolo, Matteo Negri, and Marco Turchi. 2022. Post-editing in automatic subtitling: A subtitlers' perspective. In Proceedings of the 23rd Annual Conference of the European Association for Machine Translation, pages 261–270, Ghent, Belgium, June. European Association for Machine Translation.

[Koponen et al.2020] Koponen, Maarit, Umut Sulubacak, Kaisa Vitikainen, and Jörg Tiedemann. 2020. 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, November. European Association for Machine Translation.

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

[Papineni et al.2002] Papineni, Kishore, 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, July. Association for Computational Linguistics.

[Rei et al.2020] Rei, Ricardo, 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, November. Association for Computational Linguistics.

[Sharou and Specia2022] Sharou, Khetam Al and Lucia Specia. 2022. A taxonomy and study of critical errors in machine translation. In Proceedings of the 23rd Annual Conference of the European Association for Machine Translation, pages 171–180, Ghent, Belgium, June. European Association for Machine Translation.

[Tiedemann and Scherrer2017] Tiedemann, Jörg and Yves Scherrer. 2017. Neural machine translation with extended context. In Proceedings of the Third Workshop on Discourse in Machine Translation, pages 82–92, Copenhagen, Denmark, September. Association for Computational Linguistics.

[Vincent et al.2022] Vincent, Sebastian T., Loïc Barrault, and Carolina Scarton. 2022. Controlling extra-textual attributes about dialogue participants: A case study of English-to-Polish neural machine translation. In Proceedings of the 23rd Annual Conference of the European Association for Machine Translation, pages 121–130, Ghent, Belgium, June. European Association for Machine Translation.

[Vincent et al.2023] Vincent, Sebastian, Robert Flynn, and Carolina Scarton. 2023. MTCue: Learning zero-shot control of extra-textual attributes by leveraging unstructured context in neural machine translation. In Findings of the Association for Computational Linguistics: ACL 2023, pages 8210–8226, Toronto, Canada, July. Association for Computational Linguistics.

[Vincent et al.2024] Vincent, Sebastian, Alice Dowek, Rowanne Sumner, Charlotte Blundell, Emily Preston, Chris Bayliss, Chris Oakley, and Carolina

Scarton. 2024. Reference-less analysis of context specificity in translation with personalised language models. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), Torino, Italia, May. European Language Resources Association (ELRA) and the International Committee on Computational Linguistics (ICCL).