Bayesian Hierarchical Modelling for Analysing the Effect of Speech Synthesis on Post-Editing Machine Translation

Miguel Rios, Justus Brockmann, Claudia Wiesinger,
Raluca Chereji, Alina Secară, Dragoș Ciobanu
Centre for Translation Studies, University of Vienna
{miguel.angel.rios.gaona, justus.brockmann,
claudia.wiesinger, raluca-maria.chereji,
alina.secara, dragos.ioan.ciobanu}@univie.ac.at

Abstract

Automatic speech synthesis has seen rapid development and integration in domains as diverse as accessibility services, translation, or language learning platforms. We analyse its integration in a post-editing machine translation (PEMT) environment and the effect this has on quality, productivity, and cognitive effort. We use Bayesian hierarchical modelling to analyse eye-tracking, time-tracking, and error annotation data resulting from an experiment involving 21 professional translators post-editing from English into German in a customised cloud-based CAT environment and listening to the source and/or target texts via speech synthesis. We find that using speech synthesis in the PEMT task has a non-substantial positive effect on quality, a substantial negative effect on productivity, and a substantial negative effect on the cognitive effort expended on the target text, signifying that participants need to allocate less cognitive effort to the target text.

1 Introduction

The growing adoption of data-driven approaches to machine translation (MT) since the 2000s (Kenny, 2020) has brought ongoing change to the practice of translation. While 'standard' human translation still appears to be the dominant type of service, industry surveys have repeatedly identified post-editing of MT (PEMT) as the service with the highest growth potential, according to language service providers (ELIA et al., 2023). A wealth of previous research has addressed the implications of this change, ranging from potential productivity gains (Plitt and Masselot, 2010; Läubli et al., 2019) to impacts on creativity (Guerberof-Arenas and Toral, 2022). A central theme in studies on PEMT is the effort expended by translators (Krings, 2001) and how it might be impacted by the tools they use. Moreover, previous work has probed how well PEMT is supported by the user interfaces used by translators (Moorkens and O'Brien, 2017; Herbig et al., 2020), indicating room for improvement.

A relatively novel approach to supporting PEMT processes – and translation in general – is integrating automatic text-to-speech synthesis (Taylor, 2009) in computer-assisted translation (CAT) tools. The idea is for the translator to be able to trigger an artificial voice that 'reads' to them the source and/or target text, thus adding a new mode of text reception to information processing approaches that have traditionally relied heavily on reading. Only little attention has thus far been given to this method in related work, but initial findings point to potential benefits in revision (Ciobanu et al., 2019) and PEMT (Wiesinger et al., 2022). This motivates our present study into the impact of speech synthesis on the PEMT process.

In this paper, we measure the effect of adding text-to-speech into a translation workflow for PEMT for the English-German language pair. We focus on the the target text quality delivered, cognitive effort expended, and productivity recorded, with an emphasis on the statistical modelling approach. Eye-tracking output metrics, such as the number or duration of fixations on both source and target segments are used to measure the cognitive effort during PEMT (Moorkens, 2018). Moreover, linear models and linear mixed effect models are commonly used for the analysis of eye-tracking studies (Kim et al., 2022; Silva et al., 2022), including the use of linear models to investigate the relation across text complexity of the source, cognitive effort, and PEMT (Dai and Liu, 2024). Silva et al. (2022) discuss the disadvantages of standard statistical tests for eye-tracking data analysis in subtitling. For example, t-tests conflate the data by averaging a participant into one outcome variable, and ignore other variables (predictors) that may affect the results of an experiment. Instead, they use a linear mixed-effect model for the analysis of cognitive effort (outcome variable) of reading subtitles given the effect of subtitle speeds. However, linear models require large amounts of data to achieve reliable learned estimates (Silva et al., 2022). Bayesian hierarchical models cope with data scarcity by adding information from the data structure, and prior expert knowledge that works as a regulariser to avoid over-fitting to the available data (Gelman et al., 2004). We use Bayesian hierarchical modelling to tackle the issue of data scarcity that is common in eye-tracking studies (O'Brien, 2009). Our contributions are as follows:

- We report on a PEMT study including eye-tracking, time-tracking, and quality evaluation.
- We introduce a Bayesian hierarchical model for PEMT data analysis.
- We measure how a speech-enabled mode of working may support professional translators post-editing within a CAT tool.

2 Methodology

2.1 Participants

The participants were recruited via the network of the language service provider Translated, the professional translator association UNIVERSITAS Austria, the Austrian Economic Chambers (WKO), and the website of the HAITrans research group[1]. Prospective participants were asked to fill in a recruitment questionnaire to determine whether they fulfilled the participation requirements. In total, we recruited 21 professional translators working from English into German who have German as a first language. All translators have at least three years of professional translation experience, with 10 participants having over 11 years of experience. Most participants have at least one year of PEMT experience, although five translators have little to no PEMT experience. The experiment received ethical approval from the Ethics Committee of the University of Vienna. All participants were remunerated for their time. After the conclusion of the experiment, the participant data were anonymised, and the participants were assigned an experiment ID.

2.2 Materials

The source texts used in the experiment consisted of four excerpts from two separate factsheets produced by the International Federation of Red Cross and Red Crescent Societies, UNICEF, and the World Health Organisation about stigma, mistrust, and denial in relation to COVID-19. Both factsheets were published online on the British Red Cross's Community Engagement Hub[2] in 2020.

The four English source text parts have a combined total number of 1,423 words, with their respective IDs being text 1 (t1), text 2 (t2), text 3 (t3), and text 4 (t4). To counteract the impact of the text parts on the results, we alternated text 2 and text 3 for every other participant. For this reason, we ensured comparability of the four text parts in terms of standard measurements of linguistic complexity and lexical richness as shown in Table 1, as well as readability as shown in Table 2. We use Textstat[3] for the readability scores, and LexicalRichness[4] for the linguistic complexity and lexical richness scores. The Flesch–Kincaid Reading Ease scores class all text IDs as fairly easy to read (between 80.0-70.0) and at 7th grade level. All text IDs have a consistent low linguistic complexity expressed as Type-Token Ratio (TTR).

Text	Word count	Number of syllables	Standardised TTR	Sentence count	Average sentence length
t1	342	454	0.483	18	19.0
t2	374	498	0.475	18	20.8
t3	352	471	0.520	18	19.6
t4	355	477	0.532	19	18.7

Table 1: Linguistic complexity and lexical richness for each text ID.

[1] https://haitrans.univie.ac.at/

[2] https://communityengagementhub.org/

[3] https://github.com/textstat/textstat

[4] https://github.com/lsys/lexicalrichness

Text	Flesch Reading Ease	Flesch-Kincaid Grade Level	New Dale-Chall
t1	77.57	7.2	7.58
t2	75.74	7.9	7.92
t3	76.96	7.4	7.75
t4	77.87	7.0	7.90

Table 2: Readability scores for each text ID.

2.3 Design

Before coming to the eye-tracking lab, the participants received a translation brief in German[5] with information about the task scope, target audience and style requirements, as well as the requirements for PEMT. Those five participants with little to no prior post-editing experience were also sent a short training video on MT and PEMT to watch ahead of the experiment. Upon arrival, participants signed a declaration of consent, then filled in a pre-experiment questionnaire designed to collect some demographic information and to determine their exposure to CAT tools.

The participants' task in this experiment was to post-edit the four source text parts from English into German in a customised version of the CAT tool Matecat[6] enhanced by Translated[7] with a proprietary speech synthesis function. Participants worked in two conditions: in silence, and in a sound condition whereby they could trigger speech synthesis for the source and target segments.

An EyeLink Portable Duo eye tracker[8] was used to record the participants' gaze during the experiment. Prior to performing these tasks, participants post-edited a short practice text using speech synthesis to familiarise themselves with the task setup and working environment. Each participant's computer screen and computer interactions were recorded for later annotation and comparison with other experiment participants. The total duration of the experiment was up to 3 hours.

2.4 Data Collection

The screen recordings, overlayed with participants' in-task gaze data captured with the eye tracker, were manually annotated in the SR Research Data Viewer software[9]. This included adding timestamps for task start and end times and recording the number and type of exits from the Matecat environment (e.g., to look up terms online or read the source texts made available in Microsoft Word). Areas of interest were defined around the source and target text areas in Matecat to allow for using in the analysis only the gaze data that fell within these areas.

Reports containing measures such as the total number of fixations, dwell time, and mean fixation duration for the source and target sections of the video recordings, as well as the start and end timestamps of each trial, were then generated and used for the analysis. The post-edited target texts produced by the participants were exported from Matecat for subsequent annotation and quality evaluation by multiple contributors.

When conducting eye-tracking experiments, high participant attrition rates are to be expected (O'Brien, 2009). We were able to obtain eye-tracking measures for 19 out of the 21 participants. Furthermore, due to data corruption, data from *t1* is missing entirely for one of the participants. This explains the differences in participant numbers that can be seen in Tables 3, 5, 7, and 9.

2.5 Analysis

We use Bayesian hierarchical modelling for our data analysis (Gelman et al., 2004). Hierarchical models are also known as linear mixed effects models. The motivation to use Bayesian data analysis is the data scarcity (few observations), improved learned estimates, and uncertainty quantification of the estimates. Linear regression models learn the relation of a given measurement or outcome with one or multiple predictor variables (Gelman and Hill, 2007). For example, the positive or negative effect (linear relation) of the sound condition variable on the measured quality of the produced translations.

A hierarchical model outlines a hierarchy over the data where variables are considered related or grouped under the structure of a given problem (Gelman et al., 2004). Moreover, hierarchical models take advantage of their structure to improve the learned estimates by reducing variance when the data are limited. For example, we can define groups with the produced translations by participant, condition, or type of text. A hierarchical model consists of population-level effects (fixed) for variables that describe all the observed data, and group-level effects (random) for clusters or variables that describe

[5] https://github.com/HAITrans-lab/HAITrans-bayesian-multilevel-model
[6] https://www.matecat.com/
[7] https://imminent.translated.com/
[8] https://www.sr-research.com/eyelink-portable-duo/
[9] https://www.sr-research.com/data-viewer/

variability across groups (McElreath, 2016).

Bayesian linear models allow us to test the probability of our hypothesis given the observed data by providing a posterior distribution, which contains probable values of an effect. For uncertainty quantification, Bayesian linear models produce the credible interval (CI) that is a range containing a percentage of probable values (e.g. 95%). With the given data, the effect has 95% probability of falling within this range. Moreover, Bayesian models provide a posterior distribution for the learned estimates, instead of a point from standard regression models. The posterior distribution is used to analyse the direction and size of the effect, as well as the uncertainty.

The practical importance of an effect can be decided based on the region of practical equivalence (ROPE) (Kruschke, 2018). The ROPE is a range with a small or practically no effect, which is an area that encloses values that are equivalent to the null. As a decision rule, if a large part of an estimate 95% CI falls outside from the ROPE, the effect is considered **substantial** or of **practical importance** (Kruschke, 2018). The ROPE for linear models can be defined with the standard deviation (sd) of an outcome variable as $[-0.1 * \text{sd(outcome variable)}, 0.1 * \text{sd(outcome variable)}]$.

We are interested in analysing the following outcome variables Y: Quality score based on human error annotation, Productivity with words per hour (PEMT speed), Cognitive effort with the mean fixation duration on the source text (MFD-ST), and the mean fixation duration on the target text (MFD-TT).

For the predictor variables X, we use: Condition (no sound, and sound), ID of the text (t1, t2, t3, and t4), Number of external searches, and PEMT experience (yes, no). *Condition* refers to whether the participant used speech synthesis while post-editing (sound) or not (no sound). The *text ID* identifies the text part that was post-edited. The *number of external searches* specifies how many times the participant left the CAT tool interface to perform a web search or consult other sources. *PEMT experience* refers to a participant having (yes, y) or not (no, n) previous PEMT experience.

We define a hierarchical model with random intercepts and slopes. We use the participants as the second level grouping variable to measure the effect of the sound condition on each person, and the variability across them. The population-level effects are the X predictors, and intercept and slopes for each condition and participant for group-level effects. The description of the hierarchical model is as follows:

$$
\begin{aligned}
\mathrm{y_i} &\sim \mathcal{N}\left(\mu, \sigma^2\right) \\
\mu &= \alpha_{j[i]} + \beta_{1j[i]}(\text{condition}) \\
&\quad + \beta_2(\text{text}) + \beta_3(\text{n\_searches}) \\
&\quad + \beta_4(\text{PEMT\_experience}) \\
\begin{pmatrix} \alpha_j \\ \beta_{1j} \end{pmatrix} &\sim \mathcal{N}\left(\begin{pmatrix} \mu_{\alpha_j} \\ \mu_{\beta_{1j}} \end{pmatrix}, \begin{pmatrix} \sigma^2_{\alpha_j} & \rho_{\alpha_j \beta_{1j}} \\ \rho_{\beta_{1j}\alpha_j} & \sigma^2_{\beta_{1j}} \end{pmatrix} \right) \\
&\text{, for participant j = } 1, \ldots, \text{J}
\end{aligned}
$$

where y_i is the outcome variable (e.g. quality score, PEMT speed) predicted from a normal distribution (regression) with mean μ based on a hierarchical linear model and variance σ^2. For the linear model: α_i intercept and $\beta_{1...4}$ slopes with a uniform prior are population-level coefficients, α_j intercept and β_{1j} slopes are group-level coefficients with a normal prior for each participant j.

We use the brms package in R for our Bayesian analyses (Bürkner, 2017). brms provides an interface for Bayesian linear models, and hierarchical models using Stan[10]. We show the brms formulas for our hierarchical model in the Appendix A and the scripts for our experiments are available at: `https://github.com/HAITrans-lab/HAITrans-bayesian-multilevel-model`.

3 Results

3.1 Quality

To assess quality, we scored the post-edited texts using an error typology based on the Multidimensional Quality Metrics (MQM) framework (Burchardt, 2013). Two professional translators with more than three years of experience annotated the raw MT output for the four texts using the MQM typology within the CATMA annotation tool (Gius et al., 2023). These gold standard texts are labelled with all MT errors that the participants are expected to correct according to the translation brief. The annotators first labelled the texts independently of each other, and then combined their labels into the final gold standard, asking a third annotator for advice whenever they disagreed. The MQM error severities are defined with the following weights: Minor (1), Major (5), and Critical (25). To produce the quality score for each text, we counted the number of MT errors left uncorrected, as well as errors

[10]`https://cran.rstudio.com/web/packages/brms/`

newly introduced by our participants, and weighted them according to their severity. This resulted in a score between 0 and 100 for each text, where a score of 100 would mean there were no errors in the post-edited target texts.

Condition	Text	Variable	n	mean	sd
nos	t1	quality score	20	94.635	2.469
nos	t2	quality score	10	81.311	8.054
nos	t3	quality score	11	93.828	4.672
s	t2	quality score	11	86.922	7.839
s	t3	quality score	10	88.75	4.896
s	t4	quality score	21	94.271	2.528

Table 3: Summary statistics of the *quality score* with mean and standard deviation (sd).

Population-Level Effects			
Predictors	Estimate	CI (95%)	ROPE ↓
Intercept	96.85	[92.66, 101.20]	**0.00%**
condition [s]	0.36	[-2.12, 2.82]	41.00%
text [t2]	-10.64	[-13.41, -7.89]	**0.00%**
text [t3]	-3.51	[-6.22, -0.79]	**0.00%**
text [t4]	-0.66	[-4.09, 2.76]	30.00%
n searches	-0.13	[-0.51, 0.24]	100%
PEMT experience [y]	-2.16	[-6.55, 2.29]	15.00%
Group-Level Effects			
	sd	CI (95%)	
Intercept	3.85	[2.29, 5.85]	
condition [s]	1.00	[0.04, 2.82]	

Table 4: Summary of the fitted model for the *quality score*. ROPE size ± 0.66.

Table 3 shows summary statistics with the number of participants (n), the mean, and sd of the quality score. We show the statistics grouped by both condition *no sound (nos)* and *sound (s)*, and the ID of the text (t1, t2, t3, t4).

Table 4 shows the model summary for the quality score. The predictors for the population-level effects are summarised with estimate (learned mean), 95% credible interval (CI), and percentage of the estimate that overlaps with the ROPE. The linear model takes a class or name of a variable in alphabetical order as the reference for the Intercept and adds the value of the names left as the slopes. For example, the intercept is the no sound condition *nos* and the sound condition *s* is represented with the slope condition (s).

The sound condition has a non-substantial positive effect on the quality score, because the estimate 95% CI has a large overlap with the ROPE (41%). To visualise the overlap of the sound condition CI with the ROPE, we refer the reader to Figure 6 in the Appendix. The texts t2, t3 have the highest substantial negative effect on the quality score. The effect of the number of searches (n searches) and having PEMT experience (y) are non-substantial. The group-level effect indicates how the condition (s) estimate varies from participant (group) to participant based on the sd.

To visualise the learned estimates, we show the conditional effects in Figure 1. The conditional effect plot shows the effects of each categorical or continuous predictor with the CI bar around the estimate on the outcome variable. In Figure 1 a) there is a large overlap between the CIs of the no-sound and sound conditions that indicates high uncertainty, and no difference between them. For Figure 1 b) the overlap for t2 between texts is little and indicates low uncertainty. Next, in Figure 1 c), a high number of external searches decreases the quality, but the uncertainty of the estimate is high. Moreover, in Figure 1 d), having PEMT experience (y) decreases the quality, but the difference compared to not having experience (n) is uncertain.

Figure 5 a) (Appendix) shows the fitted curve with the data points across texts from the quality score model. The posterior predictions plot shows the posterior mean (fit curve) and 95% credible interval (uncertainty bars) for each data point from the model. In other words, it plots the relation between each condition and the quality score. We can observe a difference in quality for t2, and under the sound condition, but it is small given the CI overlap.

3.2 Productivity

PEMT speed captures the number of words post-edited per hour as a measure of productivity. It was obtained by dividing the words edited (length of the respective text) by the time elapsed (task time) and then converting the result to per-hour values. Table 5 shows the summary statistics of the PEMT speed.

Condition	Text	Variable	n	mean	sd
nos	t1	PEMT speed	20	940.853	282.471
nos	t2	PEMT speed	10	1201.389	369.765
nos	t3	PEMT speed	11	853.898	220.689
s	t2	PEMT speed	11	729.944	188.497
s	t3	PEMT speed	10	1040.567	393.146
s	t4	PEMT speed	21	861.932	286.407

Table 5: Summary statistics of the *PEMT speed* with mean and standard deviation (sd).

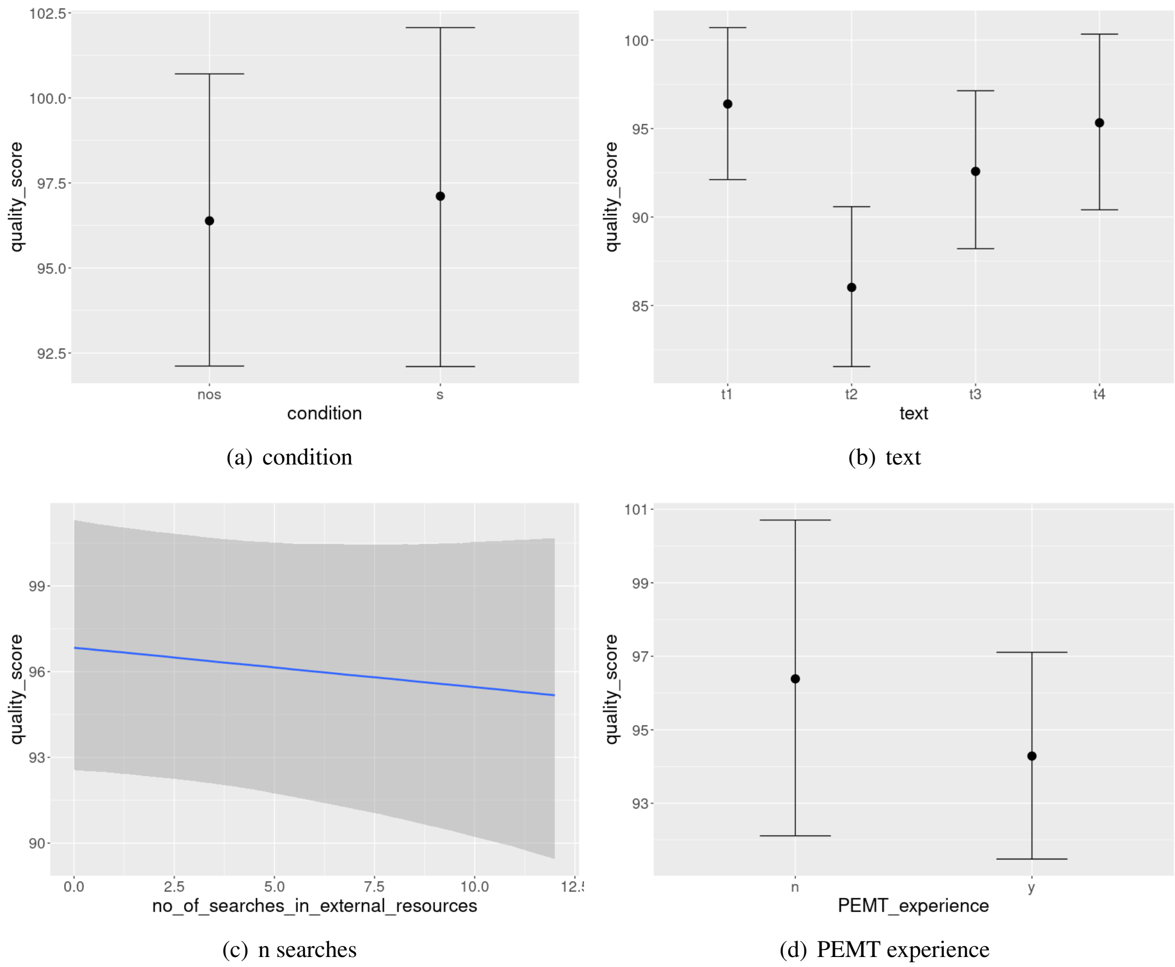


Figure 1: Conditional effects of a) condition, b) text, c) n searches, and d) PEMT experience predictors on *quality score*.

Population-Level Effects			
Predictors	Estimate	CI (95%)	ROPE ↓
Intercept	776.34	[514.94, 1031.95]	**0.00%**
condition [s]	-137.04	[-201.28, -66.16]	**0.00%**
text [t2]	88.21	[16.00, 159.16]	**3.74%**
text [t3]	64.27	[-5.62, 143.43]	15.83%
text [t4]	71.97	[-15.46, 159.45]	15.73%
n searches	-13.01	[-24.74, -1.42]	100%
PEMT experience [y]	254.74	[-34.69, 547.15]	**3.92%**
Group-Level Effects			
	sd	CI (95%)	
Intercept	287.35	[205.61, 404.32]	
condition [s]	53.27	[2.36, 129.99]	

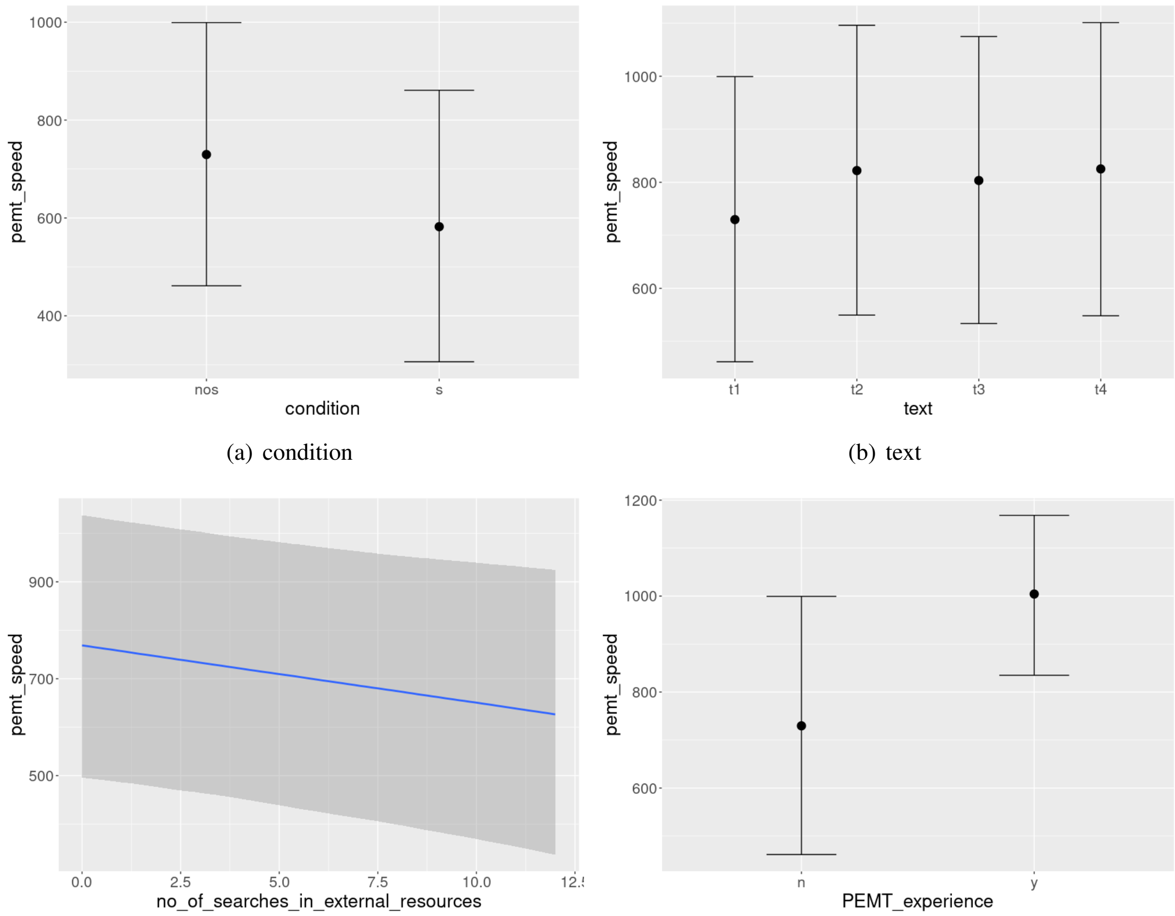
Table 6: Summary of the fitted model for the *PEMT speed*. ROPE size ±31.39.

Table 6 shows the model summary for PEMT speed with a substantial negative effect of the sound condition on the PEMT speed. There are differences across the 4 texts, with a substantial effect observed for t2. The PEMT experience has a substantial positive effect on productivity.

Figure 2 shows the conditional effects for the PEMT speed. The sound condition decreases PEMT speed in a), there is a large difference across texts in b) with the highest in t2, and an increase in the number of searches decreases the PEMT speed with high uncertainty, in c). As shown in Figure 2 d) having PEMT experience (y) increases productivity, where the difference from no experience (n) has low uncertainty. Figure 5 b) (Appendix) shows the fitted curve with the data points across texts from the productivity model. For t2 the sound condition decreases the PEMT speed, but with t3 there is an increase in speed.

3.3 Cognitive Effort

We define outcome variables for the cognitive effort with the following eye-tracking measures: MFD-ST and MFD-TT. These measures are used as a secondary indicator of the cognitive resources ex-



(a) condition (b) text

(c) n searches (d) PEMT experience

Figure 2: Conditional effects of a) condition, b) text, c) n searches, and d) PEMT experience predictors on *PEMT speed*.

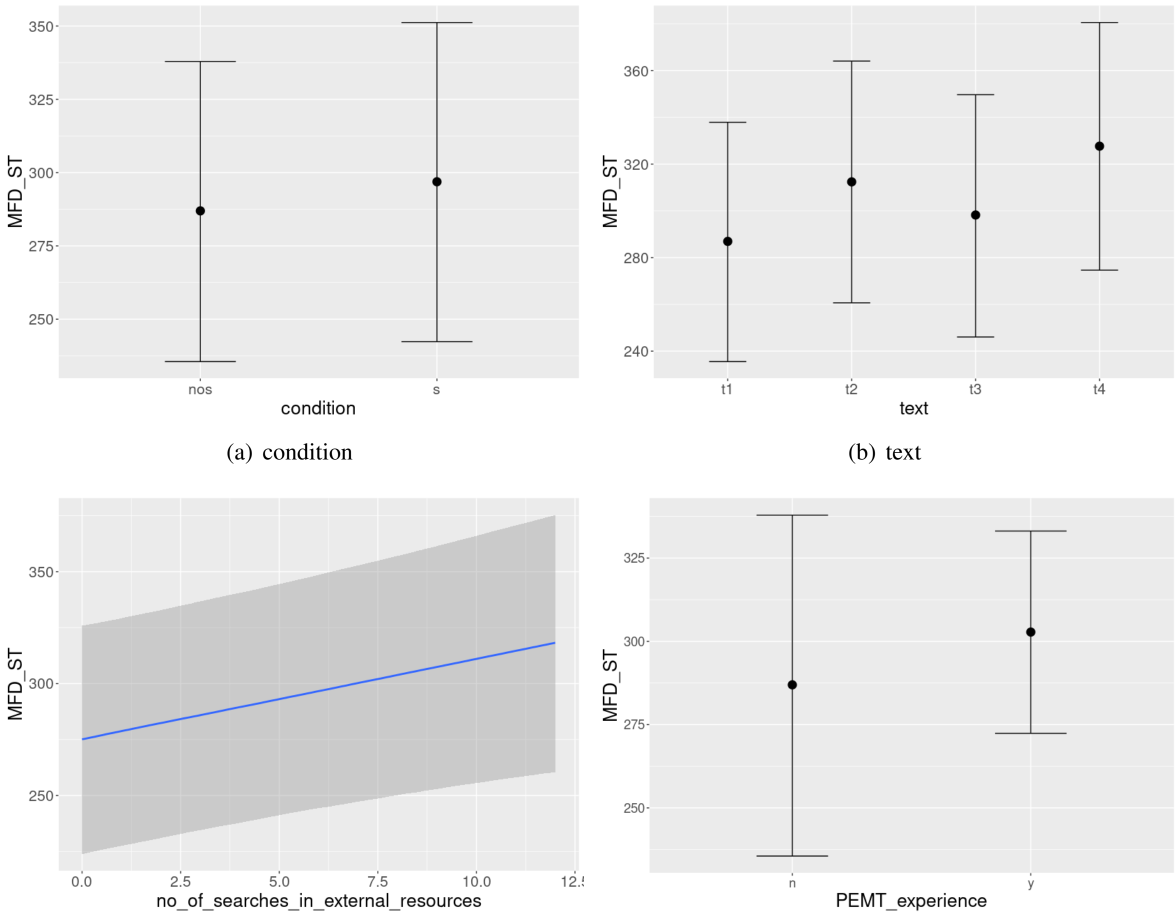
pended by participants, based on the eye-mind assumption (Just and Carpenter, 1980). Mean fixation duration is defined as the total time spent in fixations (keeping the eye stable above a point of focus), divided by the total number of fixations, and is therefore an indication of how long elements of the source and target text were fixated on average. Longer fixations are assumed to indicate higher cognitive effort (Holmqvist and Andersson, 2017). When using a method based on *visual* allocation of attention in an experiment including a listening component, it is important to note that MFD does not reflect how much time the participants spend looking at the screen, which could be assumed to be lower when adding speech synthesis to the process. Rather, MFD reflects how long fixations last on average and is therefore indicative of how effortful processing the text was for participants when they were reading it. Table 7 shows the summary statistics of the MFD-ST.

Condition	Text	Variable	n	mean	sd
nos	t1	MFD_ST	19	298.216	51.833
nos	t2	MFD_ST	9	319.582	61.672
nos	t3	MFD_ST	10	308.829	57.052
s	t2	MFD_ST	10	338.647	56.24
s	t3	MFD_ST	9	315.406	57.277
s	t4	MFD_ST	19	352.568	69.19

Table 7: Summary statistics of the *MFD-ST* with mean and standard deviation (sd).

Table 8 shows the model summary for the MFD-ST. The sound condition has a non-substantial positive effect on the MFD-ST. There are differences across the texts, with t2 and t4 having the highest effect on the MFD-ST.

Figure 3 shows the conditional effects for the MFD-ST. The sound condition increases the MFD-ST in a) with high uncertainty, there is no large difference across texts in b), the number of searches increases the MFD-ST with high uncertainty in c), and having PEMT experience (y) increases the MFD-ST with low uncertainty.



(a) condition (b) text

(c) n searches (d) PEMT experience

Figure 3: Conditional effects of a) condition, b) text, c) n searches, and d) PEMT experience predictors on *MFD-ST*.

Population-Level Effects			
Predictors	Estimate	CI (95%)	ROPE ↓
Intercept	275.10	[223.84, 325.84]	**0.00%**
condition [s]	9.85	[-7.52, 27.04]	31.61%
text [t2]	25.49	[9.34, 41.36]	**0.00%**
text [t3]	11.24	[-4.63, 26.97]	25.37%
text [t4]	40.66	[20.76, 60.30]	**0.00%**
n searches	3.61	[0.94, 6.22]	99.63%
PEMT experience [y]	16.08	[-42.96, 75.15]	14.78%
Group-Level Effects			
	sd	CI (95%)	
Intercept	55.13	[38.11, 79.49]	
condition [s]	20.45	[3.22, 38.36]	

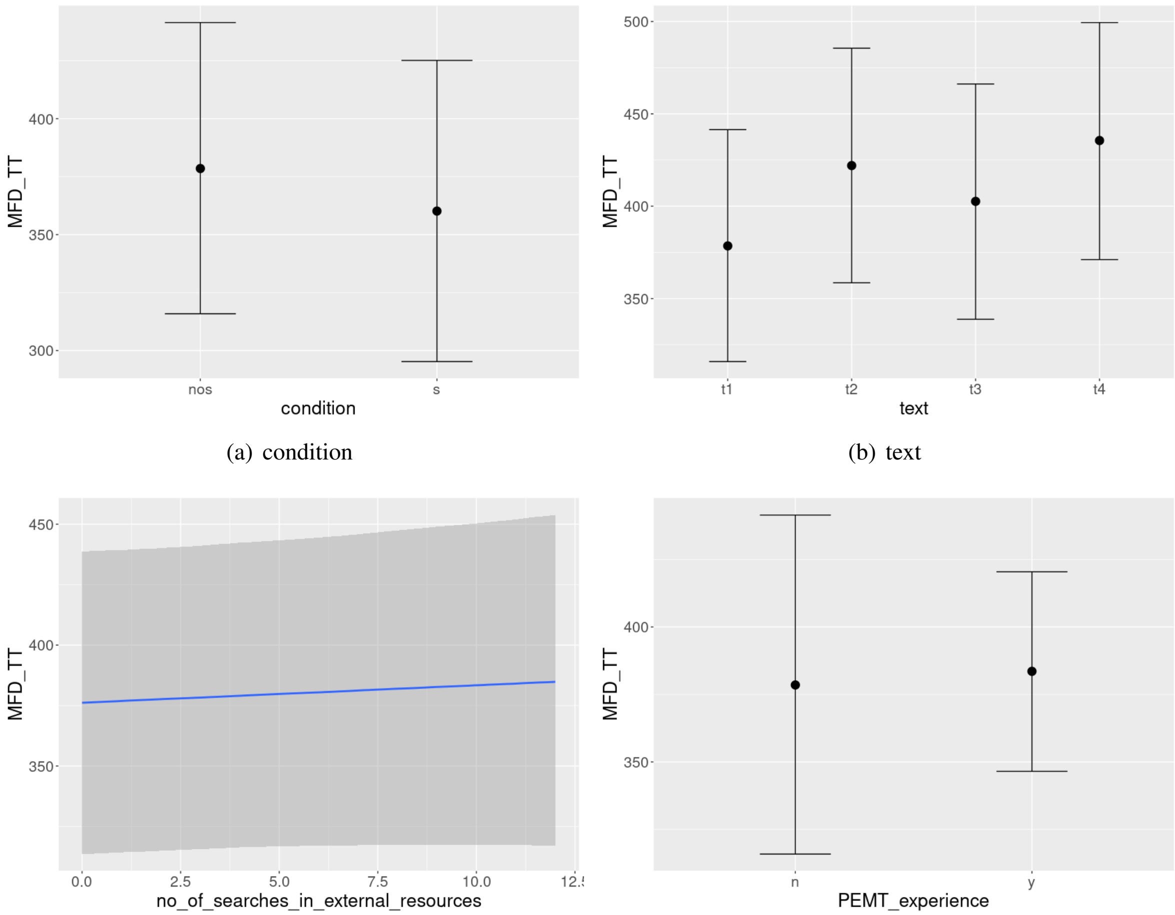
Table 8: Summary of the fitted model for the *MFD-ST*. ROPE size ±6.14.

Table 9 shows the summary statistics of the MFD-TT. Table 10 shows the model summary for the MFD-TT. The sound condition has a substantial negative effect on the MFD-TT. There are substantial differences across the texts, with t4 having the highest effect on the MFD-TT.

Condition	Text	Variable	n	mean	sd
nos	t1	MFD_TT	19	382.189	62.845
nos	t2	MFD_TT	9	416.568	69.55
nos	t3	MFD_TT	10	413.299	69.448
s	t2	MFD_TT	10	415.418	77.357
s	t3	MFD_TT	9	378.956	63.564
s	t4	MFD_TT	19	421.6	76.602

Table 9: Summary statistics of the *MFD-TT* with mean and standard deviation (sd).

Figure 4 shows the conditional effects for the MFD-TT. The sound condition decreases the MFD-TT in a) with high uncertainty, there is no large difference across texts in b), the number of searches is associated with a small increase in MFD-TT with high uncertainty in c), and having PEMT experience (y) decreases the MFD-TT in d) with low uncertainty but a large overlap with *no experience (n)*. Figure 5 (Appendix) shows the fitted curve with the



(a) condition

(b) text

(c) n searches

(d) PEMT experience

Figure 4: Conditional effects of a) condition, b) text, c) n searches, and d) PEMT experience predictors on *MFD-TT*.

Population-Level Effects			
Predictors	Estimate	CI (95%)	ROPE ↓
Intercept	376.07	[313.86, 438.67]	**0.00%**
condition [s]	-18.34	[-32.42, -4.19]	**3.41%**
text [t2]	43.28	[28.45, 58.14]	**0.00%**
text [t3]	23.96	[9.32, 38.68]	**0.00%**
text [t4]	56.95	[38.70, 75.45]	**0.00%**
n searches	0.73	[-1.74, 3.17]	100%
PEMT experience [y]	5.03	[-69.48, 79.38]	16.25%
Group-Level Effects			
	sd	CI (95%)	
Intercept	68.22	[48.20, 97.29]	
condition [s]	9.74	[0.61, 22.42]	

Table 10: Summary of the fitted model for the *MFD-TT*. ROPE size ±7.02.

data points across texts from c) MFD-ST, and d) MFD-TT. Figure c) shows that the sound condition increases the MFD-ST for t2, but decreases it for t3. The same pattern is observed for MFD-TT in d), where the sound condition is associated with an increase for t2, and a decrease for t3.

4 Discussion

The results of our experiment on using speech synthesis for PEMT indicate that (1) differences in quality between conditions were small; (2) participants were slower when using speech synthesis; and (3) participants expended less cognitive effort in TT when using speech synthesis, as reflected in their fixation data. More specifically, the presence of speech had a substantial negative effect on the MFD-TT, meaning that overall the cognitive effort spent by translators reading the target text was reduced. This may mean that hearing the target text was considered by translators to be a reliable source of information when checking PEMT. We report a non-substantial positive effect on the MFD-ST variable, indicating that the processing of the source text does not change much and only increases slightly. We do not believe this to be due to a lack of trust in the speech synthesis, given the results for the MFD-TT, but that speech use may be more worthwhile in the TT. This is also suggested by the answer to the perception questionnaire we distributed at the end of the experiment (Ciobanu et al., forthcoming) where the most reported on advantages of using speech were improved style (11/21) and error detection (9/21). It may also be that listening to the TT causes the translators to expend more cognitive effort on the ST, but this would require a separate analysis. The decrease in productivity might reflect the fact that listening to the text is an additional step to be carried out in the workflow. Moreover, as all participants but one were first-time users of speech synthesis in PEMT, productivity losses can reasonably be expected to diminish as users become more familiar with the tool. A longitudinal study would surely provide useful data in this regard. Related to this but apart from the effect of the sound condition, we also found that PEMT experience has a substantial positive effect on PEMT speed, indicating that translators with previous PEMT experience work faster than those without. The effect of the number of searches is non-substantial for all outcome variables. We recorded no substantial change in quality, but there is a perceived improvement in style and error detection for some of the participants as reported in (Ciobanu et al., forthcoming). The loss in productivity may be reduced following longer exposure to speech synthesis. This, coupled with the substantial decrease in cognitive effort in the TT, point to a potential support that a speech-enabled mode of working can offer translators.

5 Conclusions and Future Work

We quantified the impact of text-to-speech on PEMT for the English-German language pair. We introduce a Bayesian hierarchical model to tackle issues with data scarcity. The introduction of the sound condition on the PEMT workflow has a non-substantial positive effect on quality, a substantial negative effect on PEMT speed, and a non-substantial positive effect on MFD-ST and substantial negative effect on the MFD-TT for cognitive effort. The effect of the number of searches is non-substantial for all outcome variables. The text ID together with the sound condition has an effect on all of the measurements, which may be explained by the standard measurements of text complexity we used, which do not take into account semantics and might not sufficiently reflect textual differences, especially regarding translation difficulty.

For future work, we will measure the relation between text complexity evaluated with newer readability formulas based on fine-grained linguistic features and translation quality/productivity (Dai and Liu, 2024), investigate in more detail the relation between translation experience and translation quality/productivity, the relation between productivity and the number of searches performed, and quantify the observable changes in individual PEMT workflows created by our participants' access to speech synthesis.

Acknowledgements

This project received funding from the Imminent Research Grants programme. We also thank the project participants and the reviewers for their feedback.

References

Burchardt, Aljoscha. 2013. Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK, November 28-29. Aslib.

Bürkner, Paul-Christian. 2017. brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1):1–28.

Ciobanu, Dragoş, Valentina Ragni, and Alina Secară. 2019. Speech Synthesis in the Translation Revision Process: Evidence from Error Analysis, Questionnaire, and Eye-Tracking. *Informatics*, 6(4)(51), December.

Ciobanu, Dragoş, Miguel Rios, Alina Secară, Justus Brockmann, Raluca-Maria Chereji, and Claudia Wiesinger. forthcoming. The impact of speech synthesis on cognitive effort, productivity, quality, and perceptions during post-editing machine translation (PEMT). *Revista Tradumàtica: translation technologies*.

Dai, Guangrong and Siqi Liu. 2024. Towards predicting post-editing effort with source text readability: An investigation for english-chinese machine translation. *The Journal of Specialised Translation*, (41):206–229, Jan.

ELIA, EMT, EUATC, FIT EUROPE, GALA, LIND, and Women in Localization. 2023. 2023 European Language Industry Survey. Trends, expectations and concerns of the European language industry. Technical report.

Gelman, Andrew and Jennifer Hill. 2007. *Data analysis using regression and multilevel/hierarchical models*, volume Analytical methods for social research. Cambridge University Press, New York.

Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin. 2004. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd ed. edition.

Gius, Evelyn, Jan Christoph Meister, Malte Meister, Marco Petris, Mareike Schumacher, and Dominik Gerstorfer. 2023. *CATMA 7 (Version 7.0)*. Zenodo.

Guerberof-Arenas, Ana and Antonio Toral. 2022. Creativity in translation: machine translation as a constraint for literary texts. *Translation Spaces*, 11(2):184–212, November. Publisher: John Benjamins Publishers.

Herbig, Nico, Tim Düwel, Santanu Pal, Kalliopi Meladaki, Mahsa Monshizadeh, Antonio Krüger, and Josef van Genabith. 2020. MMPE: A Multi-Modal Interface for Post-Editing Machine Translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1691–1702, Online, July. Association for Computational Linguistics.

Holmqvist, Kenneth and Richard Andersson. 2017. *Eye tracking: a comprehensive guide to methods, paradigms and measures*. Lund Eye-Tracking Research Institute, Lund, Sweden, 2nd edition edition.

Just, Marcel A. and Patricia A. Carpenter. 1980. A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87(4):329–354, July. Publisher: American Psychological Association.

Kenny, Dorothy. 2020. Machine Translation. In Baker, Mona and Gabriela Saldanha, editors, *Routledge Encyclopedia of Translation Studies*, Routledge Handbooks in Translation and Interpreting Studies, pages 305–310. Routledge, 3 edition.

Kim, Yu Yeon, Aluko Ademola, Jeong Hyeun Ko, and Hee Sook Kim. 2022. Knuir at the ntcir-16 rcir: Predicting comprehension level using regression models based on eye-tracking metadata.

Krings, Hans P. 2001. *Repairing texts: empirical investigations of machine translation post-editing processes*. The Kent State University Press, Ohio.

Kruschke, John K. 2018. Rejecting or accepting parameter values in bayesian estimation. *Advances in Methods and Practices in Psychological Science*, 1(2):270–280.

Läubli, Samuel, Chantal Amrhein, Patrick Düggelin, Beatriz Gonzalez, Alena Zwahlen, and Martin Volk. 2019. Post-editing Productivity with Neural Machine Translation: An Empirical Assessment of Speed and Quality in the Banking and Finance Domain. *arXiv:1906.01685 [cs]*, June. arXiv: 1906.01685.

McElreath, Richard. 2016. *Statistical rethinking: a Bayesian course with examples in R and Stan*. Number 122 in Chapman & Hall/CRC texts in statistical science series. CRC Press/Taylor & Francis Group, Boca Raton. largely / videos.

Moorkens, Joss and Sharon O'Brien. 2017. Assessing user interface needs of post-editors of machine translation. In Kenny, Dorothy, editor, *Human Issues in Translation Technology*, pages 109–130. Routledge, London.

Moorkens, Joss, 2018. *Eye tracking as a measure of cognitive effort for post-editing of machine translation*, page 55–70. John Benjamins Publishing Company, September.

O'Brien, Sharon. 2009. Eye-tracking in translation process research: Methodological challenges and solutions. In Mees, Inger M., Susanne Göpferich, and Fabio Alves, editors, *Methodology, Technology and Innovation in Translation Process Research: A Tribute to Arnt Lykke Jakobsen*, pages 251–266. Samfundslitteratur.

Plitt, Mirko and François Masselot. 2010. A Productivity Test of Statistical Machine Translation Post-Editing in a Typical Localisation Context. *The Prague Bulletin of Mathematical Linguistics*, 93:7. Num Pages: 7 Place: Prague, Poland Publisher: De Gruyter Poland.

Silva, Breno B., David Orrego-Carmona, and Agnieszka Szarkowska. 2022. Using linear mixed models to analyze data from eye-tracking research on subtitling. *Translation Spaces*, June. © John Benjamins Publishing Company.

Taylor, Paul. 2009. *Text-to-Speech Synthesis*. Cambridge University Press.

Wiesinger, Claudia, Justus Brockmann, Alina Secară, and Dragoş Ciobanu. 2022. Speech-enabled machine translation post-editing in the context of translator training. In Kornacki, Michał and Gary Massey, editors, *Contextuality in Translation and Interpreting. Selected Papers from the Łódź-ZHAW Duo Colloquium on Translation and Meaning 2020–2021*, volume 70 of *Łódź Studies in Language*. Peter Lang.

A Model Formulas

In this section, we show the brms formulas for each outcome variable Y.

Quality score outcome: quality score, first level predictors: condition, text, n searches, PEMT experience, and second level predictors: condition. brms formula:

```
quality_score ~ 1 + condition + text + n_searches +
    pemt_experience + (1 + condition | participant )
```

PEMT speed productivity outcome: PEMT speed,first level predictors: condition, text, n searches, PEMT experience, and second level predictors: condition. brms formula:

```
pemt_speed ~ 1 + condition + text + n_searches + pemt_
    experience + (1 + condition | participant )
```

MFD-ST outcome: MFD-ST, first level predictors: condition, text, n searches, PEMT experience, and second level predictors: condition. brms formula:

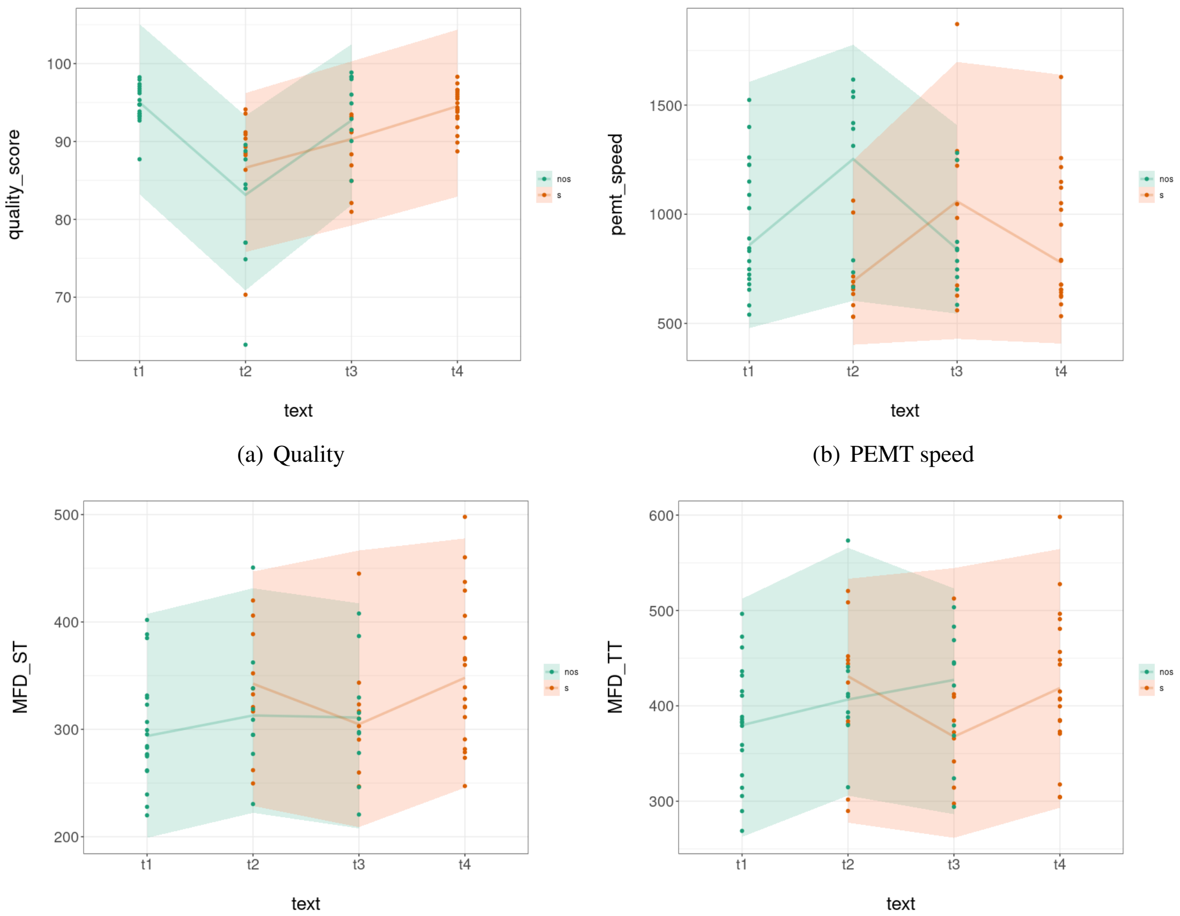
```
MFD_ST ~ 1 + condition + text + n_searches + pemt_
    experience + (1 + condition | participant )
```

MFD-TT outcome: MFD-TT, first level predictors: condition, text, n searches, PEMT experience, and second level predictors: condition. brms formula:

```
MFD_TT ~ 1 + condition + text + n_searches + pemt_
    experience + (1 + condition | participant )
```

B Fitted Models

C ROPE for the Sound Condition



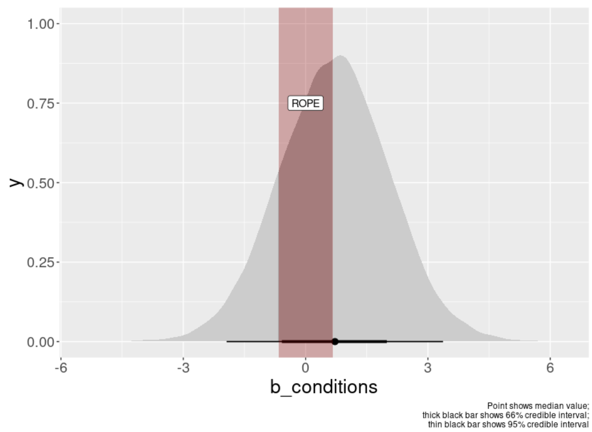
(a) Quality

(b) PEMT speed

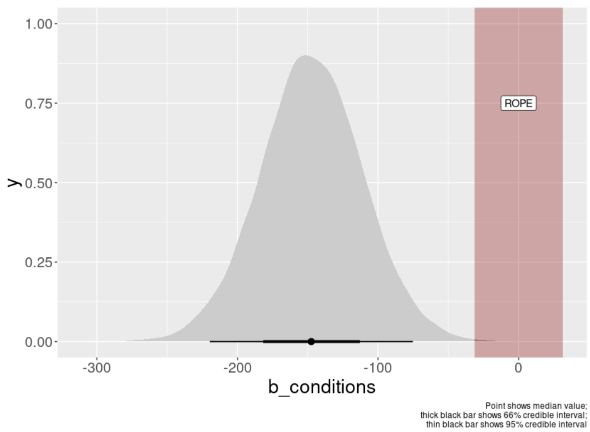
(c) MFD-ST

(d) MFD-TT

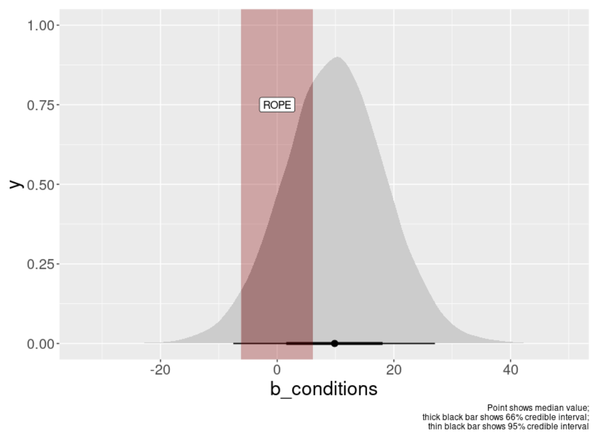
Figure 5: Fitted models across texts on each condition for: a) *Quality*, b) *PEMT speed*, c) *MFD-ST*, and d) *MFD-TT*. Fit curve with posterior predictions from the model, uncertainty bars with 95% CI, and data points.



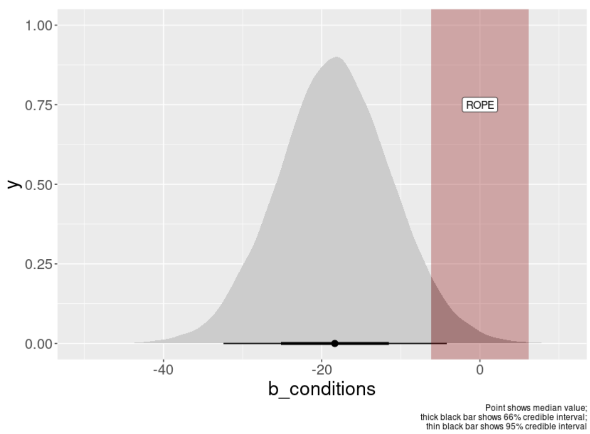
(a) quality score



(b) PEMT speed productivity



(c) MFD-ST



(d) MFD-TT

Figure 6: Proportion of the sound condition effect that falls into the ROPE for each outcome variable: a) *quality*, b) *PEMT speed*, c) *MFD-ST*, and d) *MFD-TT*. Point median value, thin bar 95% CI, and thick bar 66%CI.