Report

Imminent funded Language Research Project – Human Computer Interaction Category

The impact of Speech Synthesis on cognitive load, productivity, and quality during post-editing machine translation (PEMT)

Univ.-Prof. Dragoș Ciobanu*, Miguel Rios, Alina Secară, Justus Brockmann, Raluca-Maria Chereji, Claudia Wiesinger,

*dragos.ioan.ciobanu@univie.ac.at, miguel.angel.rios.gaona@univie.ac.at, alina.secara@univie.ac.at justus.brockmann@univie.ac.at, raluca-maria.chereji@univie.ac.at, claudia.wiesinger@univie.ac.at,

Human and Artificial Intelligence in Translation (HAITrans) research group,
University of Vienna
https://haitrans.univie.ac.at/

Overall aim: try to persuade Imminent to include synthesis
DC: check with Imminent about template/expectations

1. **Check out AFFUMT article for a bit of general story for the beginning (AS, DC)**
2. **Revisit questionnaire data (CW, RC, JB)**
    a. **Attitudes**
    b. **Willingness to use**
    c. **Interesting quotes**
3. **Keep 2 graphs with the models and replace the fancy stuff with narrative (MR)**
4. **(nice to have – maybe for EAMT) errors introduced? (CW, JB)**

Abstract

1. Introduction

Raluca

In recent years, interest in speech technologies has grown among translation researchers and industry stakeholders alike for their potential to shape translators' workflows and practices. Automatic speech recognition (ASR/dictation/speech-to-text (STT)) in particular has been shown to bring productivity and ergonomic-related improvements to translation or post-editing machine translation (PEMT) tasks, as well as an increase in output quality in scenarios where target text style and fluency are important considerations. Less attention however has been given to automatic speech synthesis

(text-to-speech (TTS)), particularly when used within and/or in conjunction with translation tools. This is partly due to misplaced fears that speech synthesis negatively impacts translator productivity by causing distractions and frustrations, despite emerging - albeit limited - evidence pointing to the contrary (refs).

For this current project, we sought to leverage the emerging interest in speech technologies and build on prior work on speech-enabled revision and PEMT tasks led by the PI at the Universities of Leeds and Vienna (refs). We investigated the impact of speech synthesis on participating translators' productivity, cognitive load and output quality and hypothesized that listening to the source and target segments when post-editing would bring improvements in Accuracy error detection compared to post-editing in silence, without also increasing cognitive load or task completion times for our participants. Our aim was to provide empirical support to the view that speech synthesis can be a viable resource in a PEMT workflow by triangulating qualitative and quantitative data collected using eye-tracking technology, time- and keystroke logging, quality analyses, as well as pre- and post-task participant questionnaires. Subsequent sections in this report provide a detailed overview of our methodology, results and findings.

2. Materials and methods

Claudia

2.1 Participants

The participants were recruited via Translated's network, the professional translator association UNIVERSITAS Austria, the Austrian Economic Chamber (WKO) and the website of the Human and Artificial Intelligence in Translation research group. Translators 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. The 5 translators who have little to no PEMT experience were required to watch a short

training video about MT and PEMT ahead of the experiment. All participants were remunerated for their time. After the conclusion of the experiment, the participant data were anonymised, and the participants were assigned a random experiment ID.

2.2 Data collection

Before coming to the eye-tracking lab, the participants received a translation brief with information about the scope, target audience and style requirements, as well as the requirements for post-editing. Upon arrival, they filled in a pre-experiment questionnaire designed to collect some demographic information and to determine their exposure to computer-assisted translation (CAT) tools and speech tools, as well as their attitudes towards post-editing and speech synthesis.

The participants' task in this experiment was to post-edit four short texts in the CAT tool Matecat, at times hearing the source and target segments being read aloud by a computer voice, and at other times without hearing any such synthetic sound. An eye tracker was used to record the participants' gaze during the experiment. Each participant's computer screen and computer interactions were recorded for later annotation and comparison with other experiment participants.

After completion of the task, the participants were asked to fill in a post-experiment questionnaire designed to capture their attitudes towards the use of speech-enabled PEMT, as well as their thoughts on the viability of the workflow and any challenges that they encountered. The total duration of the experiment was up to 3 hours.

2.3 Text preparation

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 Organization about stigma, mistrust, and denial in relation to COVID-19. Both factsheets were published online on the British Red Cross's Community Engagement Hub[1] in 2020.

The four English source text parts have a combined total number of 1,423 words. To counteract the impact of fatigue and growing familiarity with the texts, 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 linguistic complexity and lexical richness (see Table 1), as well as readability (see Table 2).

	Source word count (without punctuation)	**Number of syllables**	**Standardised type-token ratio (TTR)**	**Total sentence count**	**Average sentence length in words**
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 based on Textstat[2] and LexicalRichness[3].

	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 based on Textstat.

The machine translation output for all four text parts was retrieved via Matecat on the same day and using the same settings. The MT engine used to produce the translations was the integrated version of ModernMT.

2.4 Multilevel Models

Report

We use Bayesian multilevel (hierarchical) modelling for our data analysis [BIB]. The motivation to use Bayesian data analysis is the data scarcity (few observations), improved learned estimates, and the included uncertainty quantification. Linear regression models learn the relation of a given measurement or outcome with one or multiple predictor variables [BIB]. For example, the positive or negative effect (linear relation) of the sound condition variable on the measured quality of the produced translations.
A multilevel model outlines a hierarchy over the data where variables are considered related or grouped under the structure of a given problem [BIB]. For example, we can define groups with the produced translations by participant, condition, or type of text. The multilevel 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 variability across clusters [BIB].

We are interested in analyzing the following outcome variables *Y*: cognitive load with the mean fixation duration of the source text (MFD-ST) and the mean fixation duration of the target text (MFD-TT), the quality score, and the productivity with words per hour and translation edit rate score (TER). For the predictor variables *X*, we use the sound condition (no sound, and sound), and id of the text (*t1, t2, t3*, and *t4*). We use the participants as the second level of the model (group-level effect) to measure the effect of the sound condition on each person, and the variability across them.

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, for example 95%. With the given data the effect has 95% probability of falling within this range. We use the brms package in R for our Bayesian analyses [BIB]. Brms provides an interface for Bayesian linear models, multilevel models using Stan.

[1] https://communityengagementhub.org/
[2] https://github.com/textstat/textstat
[3] https://github.com/LSYS/LexicalRichness

3. Results

3.1 Cognitive load

Dragoș, Alina, Miguel, Raluca

We define outcome variable for cognitive load of the participants with mean fixation duration of the source text (MFD-ST) and mean fixation duration of the target text (MFD-TT)….

The number of participants differ to 18 because...

Table X. shows summary statistics with the mean and standard deviation (sd) of the MFD-ST. We show the statistics group by both conditions no sound (**nos**) and sound (**s**) and the id of the text.

Condition	Text id	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	9	308.104	60.464
s	t2	MFD_ST	9	341.091	59.085
s	t3	MFD_ST	9	315.406	57.277
s	t4	MFD_ST	19	352.568	69.19

Figure X. shows the multilevel model summary for the MDF-ST. The variables are summarised with *Estimate* the learned mean, *Est. Error* the standard deviation, 95% *credible interval* (CI), and standard deviation of a group-level variable *sd*(.). 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 positive effect on the MFD-ST. The *t2, t3,* and *t4* also have *a* positive effect. The text *t4* has the highest effect on the MFD-ST.

Report

```
 Family: gaussian
  Links: mu = identity; sigma = identity
Formula: MFD_ST ~ 1 + condition + text + (1 + condition | participant)
   Data: eyetracking (Number of observations: 76)
  Draws: 4 chains, each with iter = 10000; warmup = 1000; thin = 1;
         total post-warmup draws = 36000

Group-Level Effects:
~participant (Number of levels: 19)
                          Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)                52.61     10.04    36.57    75.70 1.00    12095    19237
sd(conditions)               17.29      8.83     1.66    35.92 1.00     8710    11203
cor(Intercept,conditions)     0.35      0.38    -0.49     0.95 1.00    21543    17529

Population-Level Effects:
           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept    297.73     13.03   272.04   323.49 1.00     8511    13551
conditions    12.71      8.92    -4.71    30.40 1.00    23666    25250
textt2        23.96      8.93     6.45    41.67 1.00    26371    26937
textt3         8.41      8.75    -8.95    25.54 1.00    27836    26698
textt4        41.51     11.03    19.84    63.26 1.00    23481    25356

Family Specific Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma    23.81      2.73    19.00    29.70 1.00    13675    22261

Draws were sampled using sample(hmc). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

Bayesian linear models provide a posterior distribution for the learned estimates, instead of a point from standard regression models [BIB]. The posterior distribution will allow us to analyse the direction and size of the effect, as well as the uncertainty. In Figure X., we show the posterior distribution of the population level effect for the sound condition (condition**s**) to highlight the size of the effect. The region of practical equivalence (ROPE) is a range with a small or practically no effect. The ROPE is defined as -0.1 * sd(outcome variable) to 0.1 * sd(outcome variable) [BIB]. If a large part of the CI falls outside of the ROPE results in a substantial effect. For the sound condition (condition**s**) 78.5% of the CI falls outside the ROPE. In other words, the positive effect of the sound condition on the MFD-ST is small, where participants...

Report

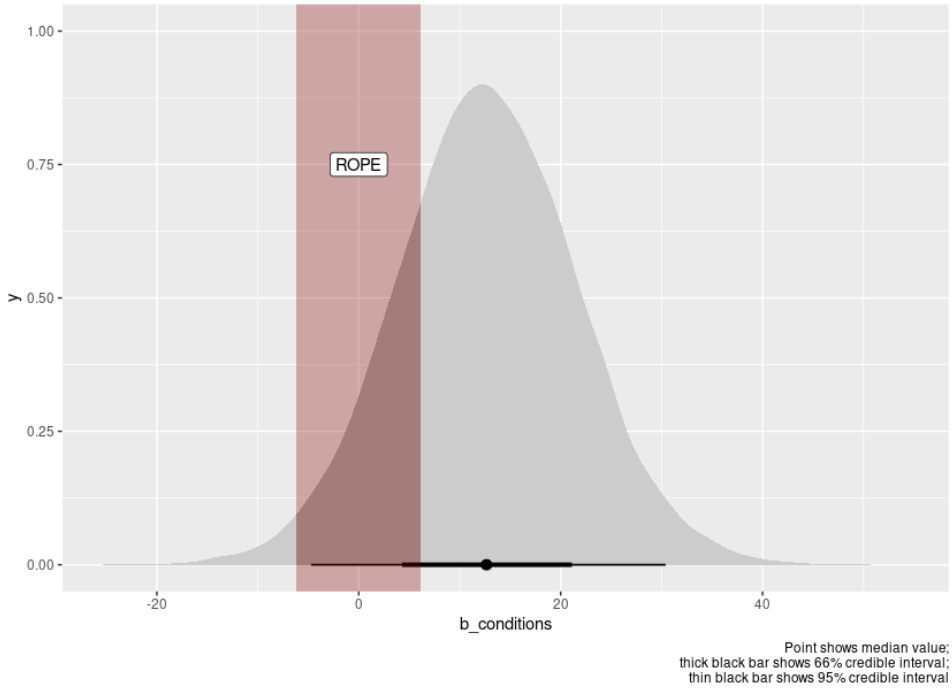
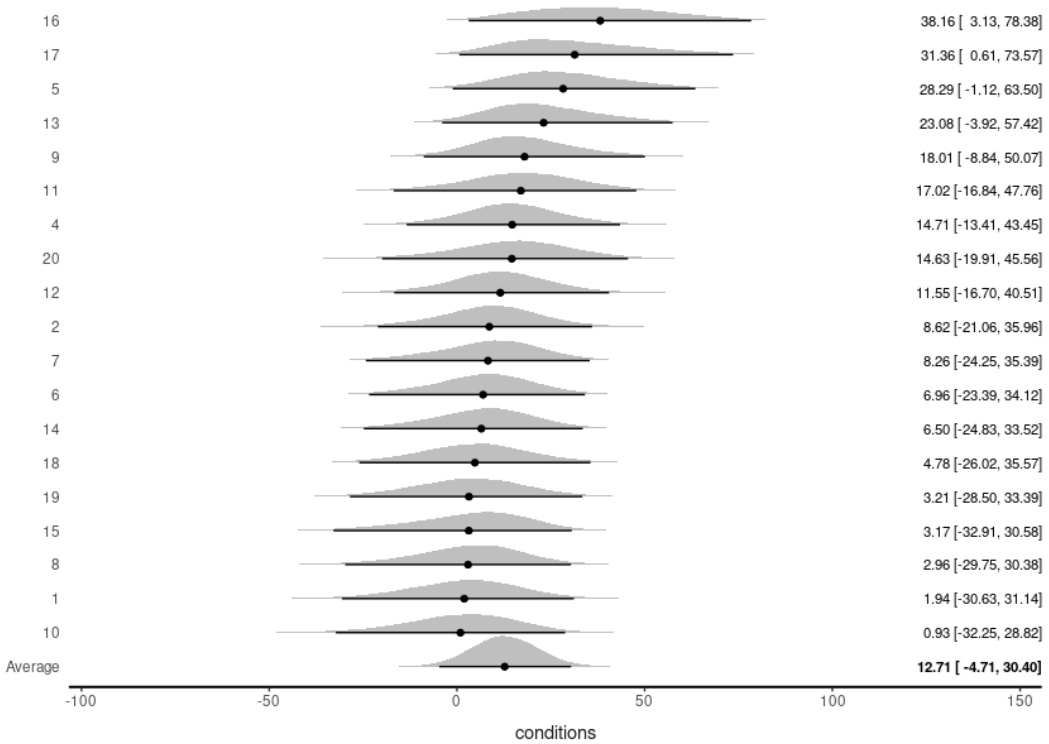


Figure X. shows the group-level effects of the sound condition across participants. The figure sorts the effects for each participant, for example, the highest effect (38.16) is on *participant 16*. However, the uncertainty (variability) is high [3.13, 78.38].



Report

Table X. shows the summary statistics of the MDF-TT.

Condition	Text id	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	9	409.854	72.749
s	t2	MFD_TT	9	414.413	81.98
s	t3	MFD_TT	9	378.956	63.564
s	t4	MFD_TT	19	421.6	76.602

For the MFD-TT, we show the multilevel model summary for the MFD-TT in Figure X. The sound condition has a negative effect on the MFD-TT, and the texts have a positive effect.

```
 Family: gaussian
  Links: mu = identity; sigma = identity
Formula: MFD_TT ~ 1 + condition + text + (1 + condition | participant)
   Data: eyetracking (Number of observations: 76)
  Draws: 4 chains, each with iter = 10000; warmup = 1000; thin = 1;
         total post-warmup draws = 36000

Group-Level Effects:
~participant (Number of levels: 19)
                          Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)                65.90     11.95    47.07    93.55 1.00     6862    13332
sd(conditions)                9.86      5.78     0.64    22.30 1.00    14912    13987
cor(Intercept,conditions)     0.53      0.41    -0.56     0.99 1.00    25372    19099

Population-Level Effects:
           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept    382.18     15.33   351.75   412.17 1.00     5661    10412
conditions   -17.69      7.13   -31.73    -3.62 1.00    20108    24501
textt2        42.82      7.55    27.89    57.65 1.00    22753    25313
textt3        23.35      7.32     8.94    37.77 1.00    22883    24570
textt4        57.03      9.37    38.57    75.34 1.00    20070    23824

Family Specific Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma    20.15      2.11    16.52    24.81 1.00    23056    23392

Draws were sampled using sample(hmc). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

Report

Figure X. shows the posterior distribution of the population level effect for the sound condition on the MFD-TT. For the sound condition 95.8% of the CI falls outside of the ROPE. The negative effect of the sound condition is **substantial**, and participants spend less time on...

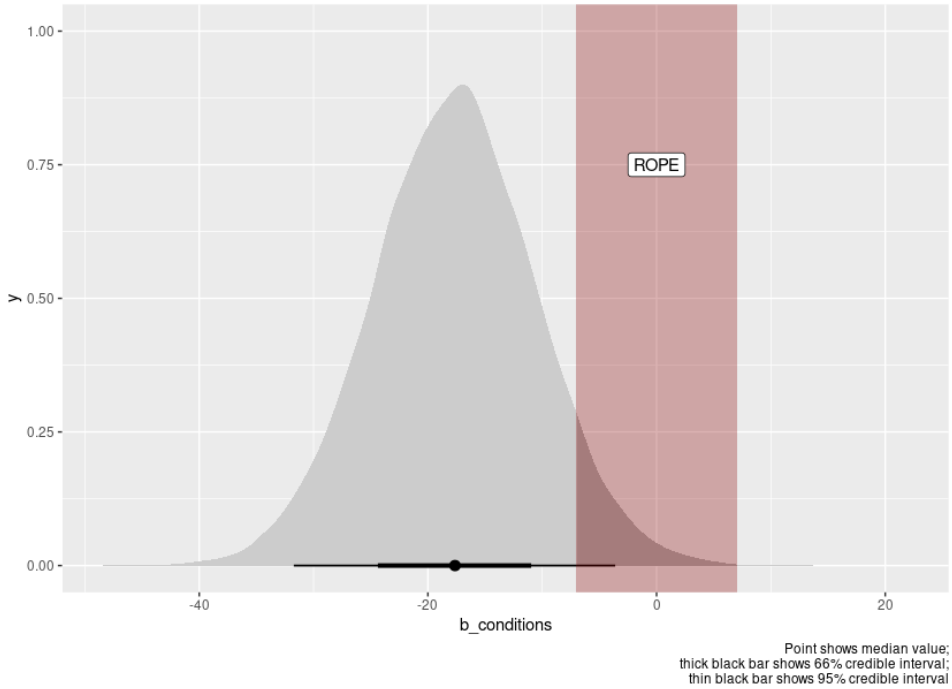
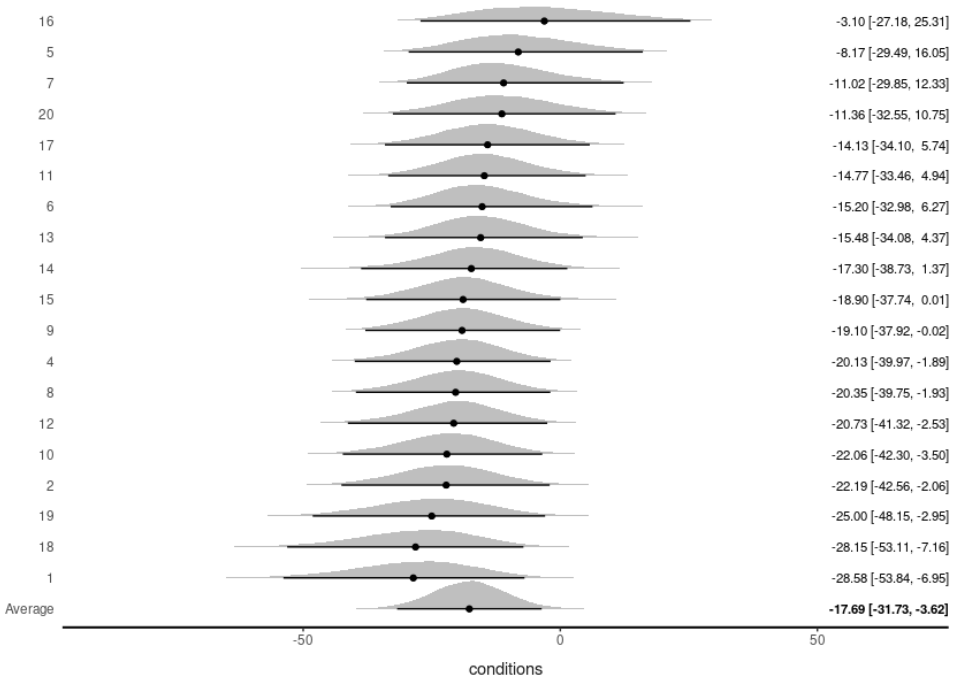


Figure X. shows the group-level effects of the sound condition across participants.

Report



3.2 Productivity

Dragoș, Alina, Miguel

Finding a reliable way to measure productivity and particularly the effort that goes into PEMT is something that many spent time investigating.

Trados Studio provides a plug-in called Post-Edit Compare which captures the changes made on an MT pretranslated file during PEMT. The user can visualise as Track-Changes the individual changes made and an Edit Distance calculation based on the Damerau-Levenshtein method is also provided. A Post Edit Modification % score is provided. Should we use PEM or the actual number of changes as we have them now under Edit distance?
Determining effort on number of words alone is problematic as time spent reading the source and the target, understanding them and making a judgement about quality, in addition to time spent consulting necessary external resources are not captured.
Post-Edit compare can also track time, but the suggestion is to use Qualitivity for that.
Matecat editdistance
Matecat QR – did we mean to export data info, too? It has the time to edit values
Tasktime eyetracking (ST+TT time+time in references)
ST+TT time
Which ones of the two above is closeer to the values reported in matecat?
In matecat Help,
“If you are a project manager, you can use the Quality Report in Matecat to keep track of quality for your projects.

Report

The score is updated in real-time and you can easily identify at a glance all flagged issues, in order to take action.
The indications about Post-Editing Effort (PEE) and Time to Edit (TTE) provide insights about productivity and the actual effort it took to complete the job.
NOTE: Time-to-edit is calculated by adding up the time a translator spends on each segment. If a translator goes back to the same segment several times (opening and closing the segment), the times for each edit are added together.
For the calculation of the Post-Editing Effort, on the other hand, some segments are excluded:

- Segments that have a time-to-edit of over 25 seconds per word, and
- Segments that have a time-to-edit below 0.5 seconds per word."

The TTE values are per segment, just like the edit Distance from Trados and exist in the exports from matecat under participant translations.

Matecat Speed: x Words/h is meant to provide an overview of the workload, more like a speed tracker, time estimated to complete a task, and depends on how quickly a segment is confirmed so the productivity below is actually not linked to productivity.

We define two models for productivity with PEMT speed (words/hour) and TER score on a text. TER is the number of edit operations that a required for a hypothesis to match a reference translation. We use TER between the hypothesis MT and the PEMT for a text.

PEMT speed simple model

Table X. shows the summary statistics for the PEMT speed.

Report

Condition	Text id	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

```
 Family: gaussian
  Links: mu = identity; sigma = identity
Formula: pemt_speed ~ 1 + condition + text + (1 + condition | participant)
   Data: eyetracking (Number of observations: 83)
  Draws: 4 chains, each with iter = 10000; warmup = 1000; thin = 1;
         total post-warmup draws = 36000

Group-Level Effects:
~participant (Number of levels: 21)
                         Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)              303.72     52.36   219.65   423.69 1.00     8352    14032
sd(conditions)              53.06     35.13     2.54   130.60 1.00     7658    13061
cor(Intercept,conditions)    0.11      0.45    -0.81     0.92 1.00    29137    18347

Population-Level Effects:
           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept    931.82     69.11   794.82  1066.09 1.00     5931    10062
conditions  -142.98     34.67  -212.12   -74.57 1.00    23432    25735
textt2        95.25     37.04    21.85   167.88 1.00    22499    25031
textt3        69.84     36.19    -1.54   140.67 1.00    24127    26009
textt4        67.24     45.21   -21.97   155.71 1.00    19919    22927

Family Specific Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma   102.32     10.76    83.10   125.29 1.00    12770    20823

Draws were sampled using sample(hmc). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

Figure X. shows the model summary with a negative effect of the sound condition on the PEMT speed, and a positive effect of the texts with highest *t2.*

In the box above: the model learns and estimates that in the ns the average speed is 931w/h and in s 789w/h (931-142). For T2, in general the speed is 931 + 95.25 T3 931+69.84

Report

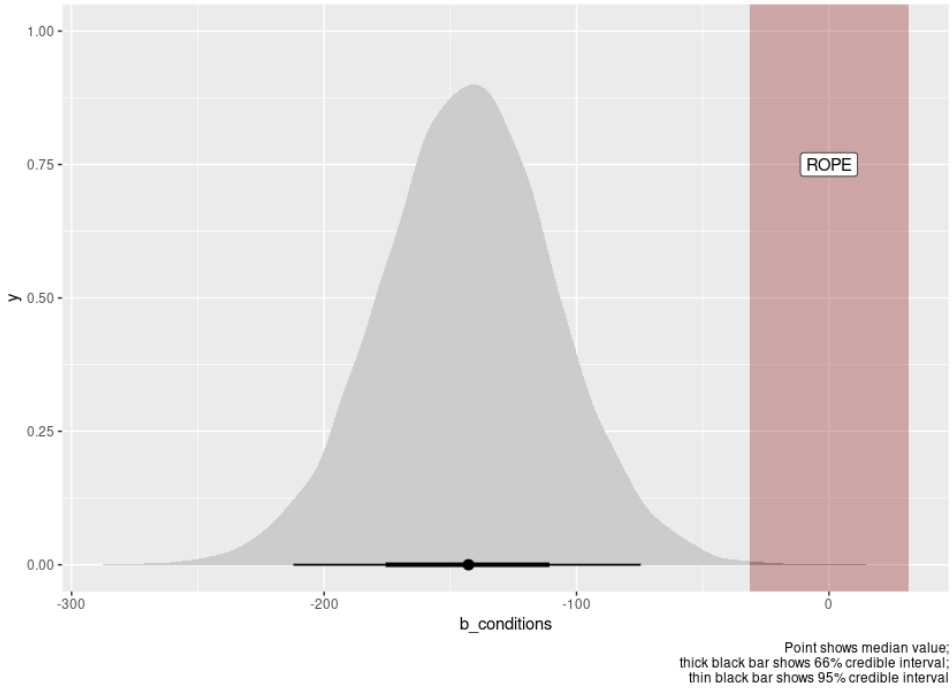
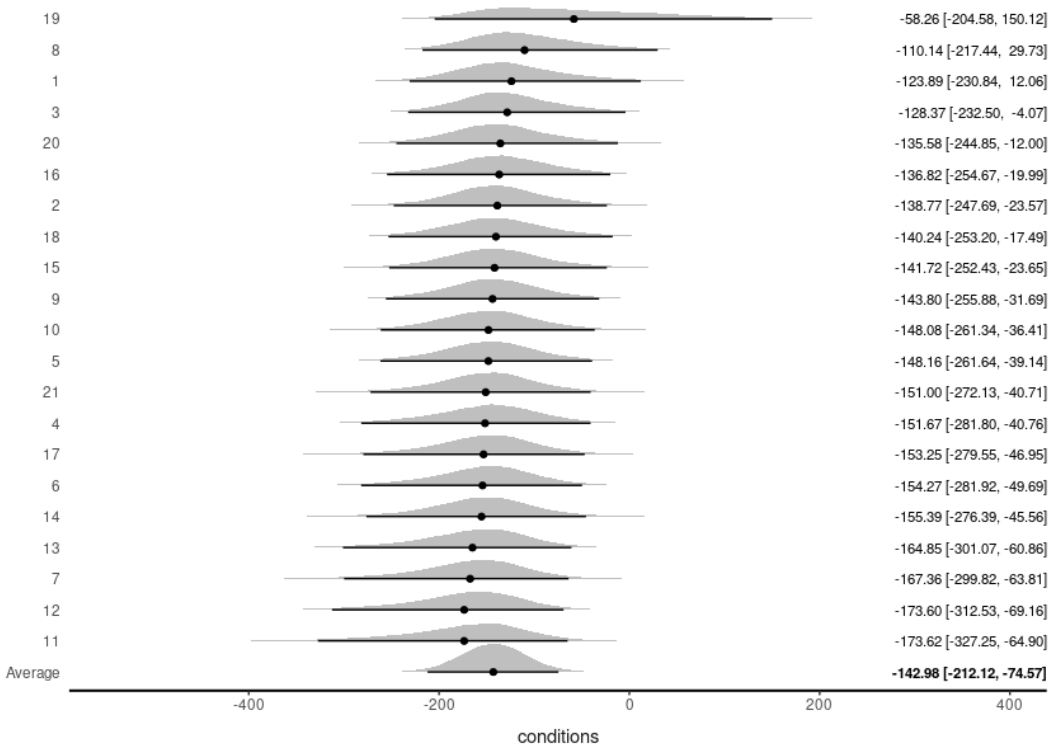


Figure X. shows the posterior distribution of the population level effect for the sound condition on the PEMT speed. For the sound condition all the CI falls outside the ROPE, and the negative effect of the sound condition is **substantial**.

If no overlap, like in the rope above, the effect is substantial.

Figure X. shows the group-level effects of the sound condition across participants for the PEMT speed.

Report



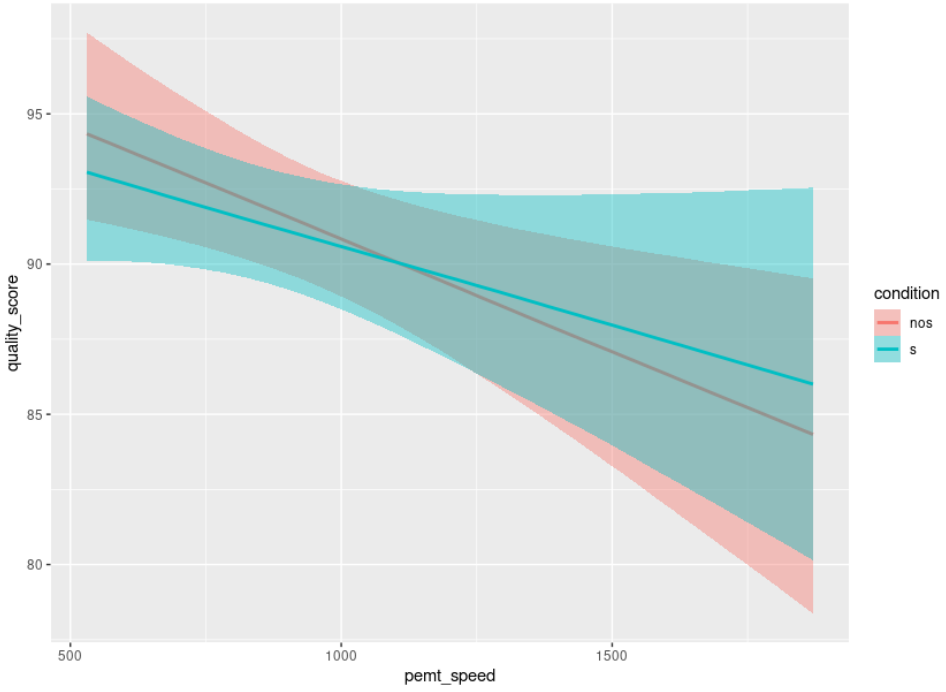
Effect of PEMT Speed on quality

In addition, we define a multilevel model with the condition as the group-level effect to measure the effect of PEMT speed group by condition over the quality score and the TER score.

Report

```
 Family: gaussian
  Links: mu = identity; sigma = identity
Formula: quality_score ~ 1 + pemt_speed + (1 + pemt_speed | condition)
   Data: eyetracking (Number of observations: 83)
  Draws: 4 chains, each with iter = 10000; warmup = 1000; thin = 1;
         total post-warmup draws = 36000

Group-Level Effects:
~condition (Number of levels: 2)
                         Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)                3.54      3.45     0.05    11.64 1.08       38       76
sd(pemt_speed)               0.17      0.43     0.00     1.47 1.50        8       24
cor(Intercept,pemt_speed)   -0.18      0.57    -0.99     0.92 1.11       26       58

Population-Level Effects:
           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept     97.92      3.89    90.75   106.68 1.07       39       66
pemt_speed    -0.03      0.05    -0.13     0.00 1.55        7       14

Family Specific Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma     6.43      0.54     5.47     7.64 1.01      148      155

Draws were sampled using sample(hmc). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

Figure X. shows the effect of the PEMT speed on quality under the sound conditions. The effect of both conditions is negative, however with a higher PEMT speed the sound condition will decrease less than with no sound.



TODO add text into model???

Report

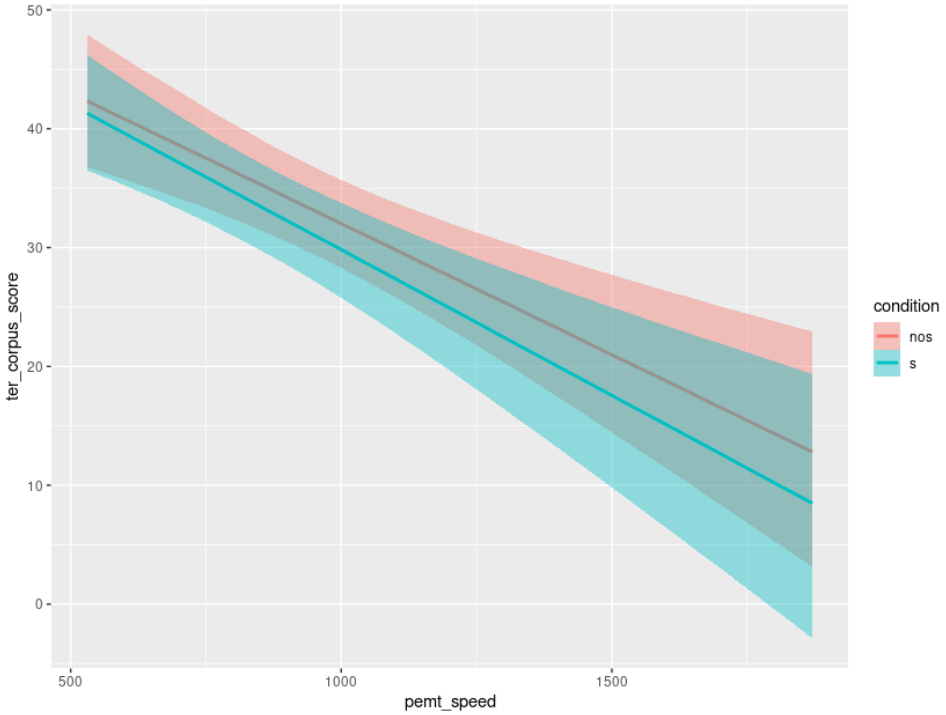
Effect of PEMT speed on TER

```
 Family: gaussian
  Links: mu = identity; sigma = identity
Formula: ter_corpus_score ~ 1 + pemt_speed + (1 + pemt_speed | condition)
   Data: eyetracking (Number of observations: 83)
  Draws: 4 chains, each with iter = 10000; warmup = 1000; thin = 1;
         total post-warmup draws = 36000

Group-Level Effects:
~condition (Number of levels: 2)
                         Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)                8.18      8.28     0.23    30.34 1.00     7577    11330
sd(pemt_speed)               0.02      0.02     0.00     0.09 1.01      808      150
cor(Intercept,pemt_speed)   -0.09      0.61    -0.98     0.95 1.01      858      401

Population-Level Effects:
           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept     54.03      7.32    38.57    68.80 1.00     7749     6272
pemt_speed    -0.02      0.01    -0.05    -0.00 1.00     3590     2305

Family Specific Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma    12.41      1.01    10.65    14.56 1.00     3217    12119

Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

Figure X. shows the model summary with a negative effect of the PEMT speed on the TER score. In other words, fewer edit operations....

Figure X. shows the conditional effect of the PEMT speed on the TER score grouped by condition. For the sound condition...

Report



TER simple model

Table X. shows the summary statistics of the TER score.

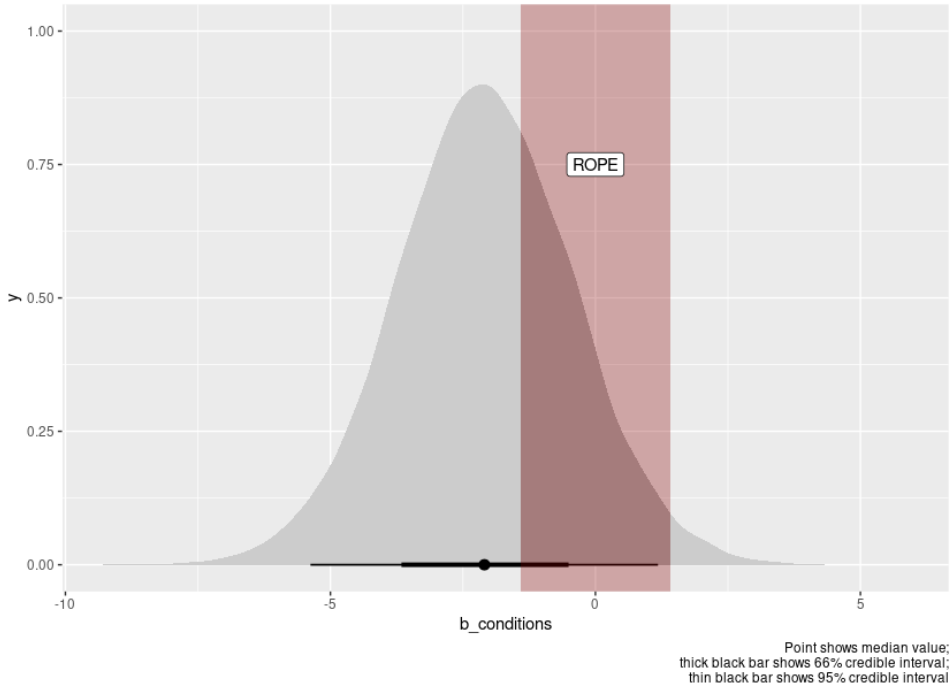
Condition	Text id	Variable	n	mean	sd
nos	t1	ter corpus score	21	30.337	14.667
nos	t2	ter corpus score	10	31.903	13.07
nos	t3	ter corpus score	11	37.438	13.509
s	t2	ter corpus score	11	39.282	15.244
s	t3	ter corpus score	10	25.879	12.627
s	t4	ter corpus score	21	33.012	13.353

Report

```
 Family: gaussian
  Links: mu = identity; sigma = identity
Formula: ter_corpus_score ~ 1 + condition + text + (1 + condition | participant)
   Data: eyetracking (Number of observations: 83)
  Draws: 4 chains, each with iter = 10000; warmup = 1000; thin = 1;
         total post-warmup draws = 36000

Group-Level Effects:
~participant (Number of levels: 21)
                          Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)                14.05      2.43    10.13    19.51 1.00     7754    14685
sd(conditions)                1.87      1.42     0.08     5.22 1.00    11653    16524
cor(Intercept,conditions)    -0.07      0.51    -0.92     0.91 1.00    35609    21634

Population-Level Effects:
           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept     30.47      3.24    24.07    36.94 1.00     5860    11184
conditions    -2.09      1.66    -5.38     1.18 1.00    26424    26363
textt2         6.44      1.83     2.85    10.05 1.00    24853    25840
textt3         2.30      1.79    -1.21     5.80 1.00    25115    27521
textt4         4.57      2.26     0.08     9.00 1.00    22084    25348

Family Specific Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma     5.10      0.51     4.20     6.20 1.00    20938    25227

Draws were sampled using sample(hmc). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

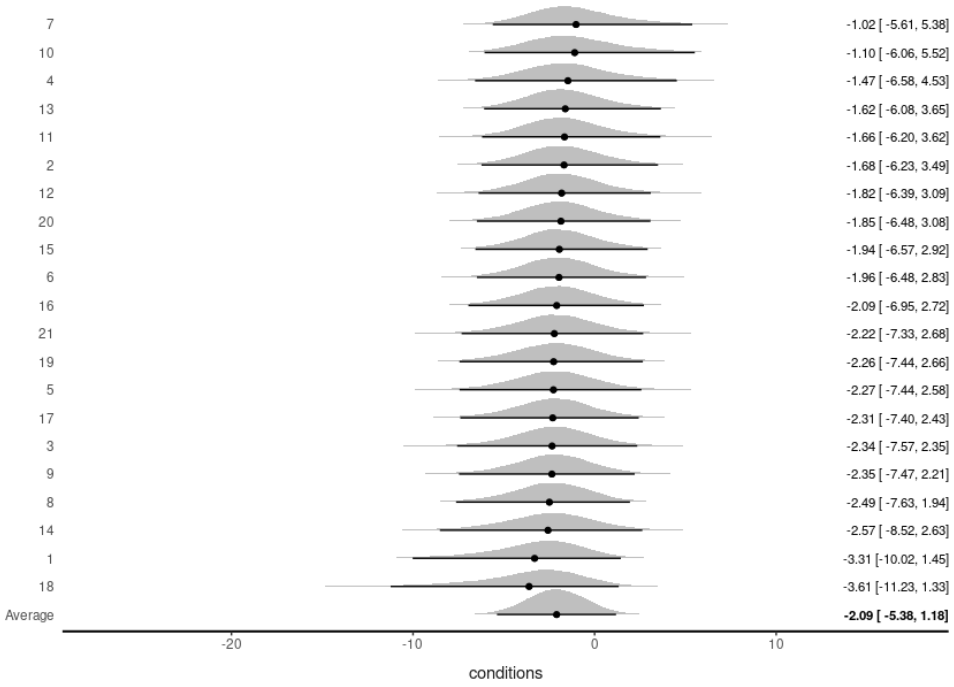
Figure X. shows the model summary for the TER score, where the sound condition has a negative effect and the texts positive. However, in Figure X. the sound condition CI falls X% in the ROPE thus this negative effect is small.

Report



As there is overlap, the effect is small.

Figure X. shows the group-level effects of the sound condition across participants for the TER score.



3.3 Quality

To be able to assess the quality of the participants' post-edits, two members of the research team with more than 3 years of translation experience created Gold Standard versions of the four text parts using an error typology based on the MQM framework [1] and tailored to the needs of the project. These Gold standard versions flag all MT errors that the participants were expected to correct in line with the translation brief. In doing so, the team members first annotated the texts independently of each other and then combined their annotations into a final Gold Standard, asking a third colleague for advice whenever they disagreed. Error severities were annotated as *Minor*, *Major*, or *Critical*.
The Gold Standard errors still present in the participants' final post-edited texts, as well as the errors introduced by the participants themselves during post-editing, were counted and weighed by their severity to arrive at a quality score for each text.

Table X. shows the quality score summary statistics.

Condition	Text id	Variable	n	mean	sd
nos	t1	quality score	21	94.723	2.44

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

Figure X. shows the multilevel model summary for the quality score. The sound condition has a positive effect on the quality, and the texts have a negative effect on the condition. The text t2 has the lowest negative effect...

```
 Family: gaussian
  Links: mu = identity; sigma = identity
Formula: quality_score ~ 1 + condition + text + (1 + condition | participant)
   Data: eyetracking (Number of observations: 84)
  Draws: 4 chains, each with iter = 10000; warmup = 1000; thin = 1;
         total post-warmup draws = 36000

Group-Level Effects:
~participant (Number of levels: 21)
                         Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)                3.70      0.87     2.18     5.61 1.00    11975    17028
sd(conditions)               0.96      0.75     0.04     2.77 1.00    15228    15765
cor(Intercept,conditions)   -0.08      0.56    -0.95     0.93 1.00    33130    22997

Population-Level Effects:
           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept     94.81      1.18    92.51    97.11 1.00    14495    21012
conditions     0.27      1.26    -2.20     2.75 1.00    24094    25842
textt2       -10.61      1.38   -13.33    -7.89 1.00    23461    25918
textt3        -3.44      1.35    -6.09    -0.76 1.00    23738    26326
textt4        -0.71      1.73    -4.11     2.71 1.00    19974    23493

Family Specific Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma     3.95      0.38     3.29     4.79 1.00    22498    23880

Draws were sampled using sample(hmc). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

Figure X. shows the posterior distribution of the population level effect for the sound condition with 55.7% of the CI falls outside of the ROPE, thus the negative effect of the sound condition is small.

Report

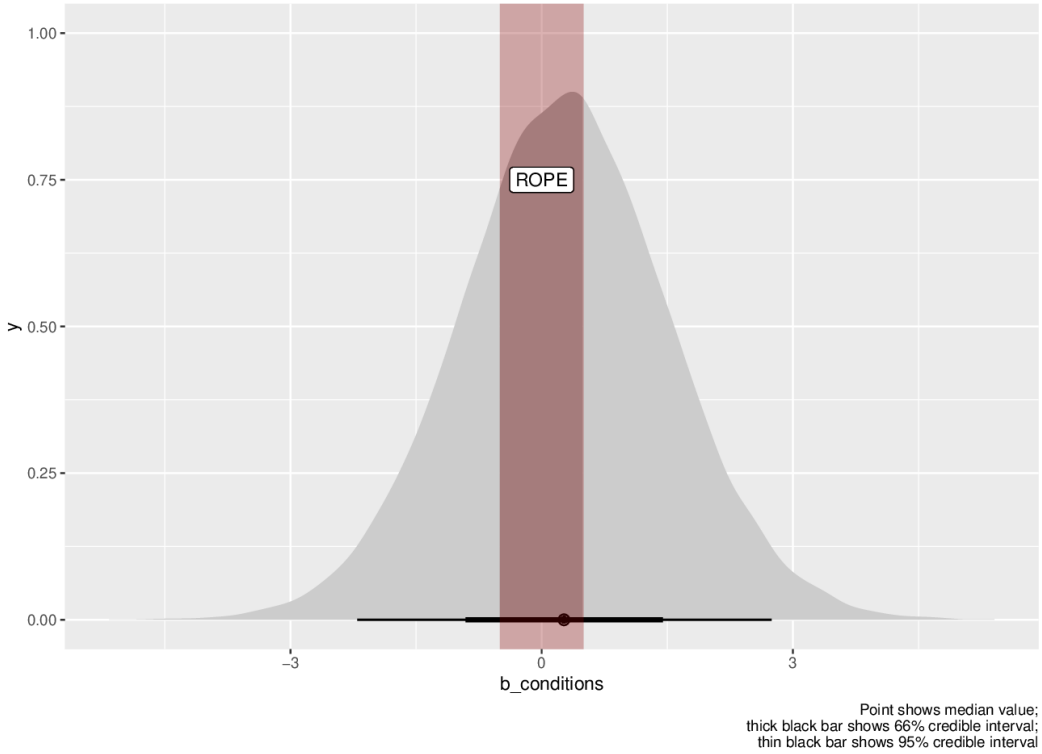
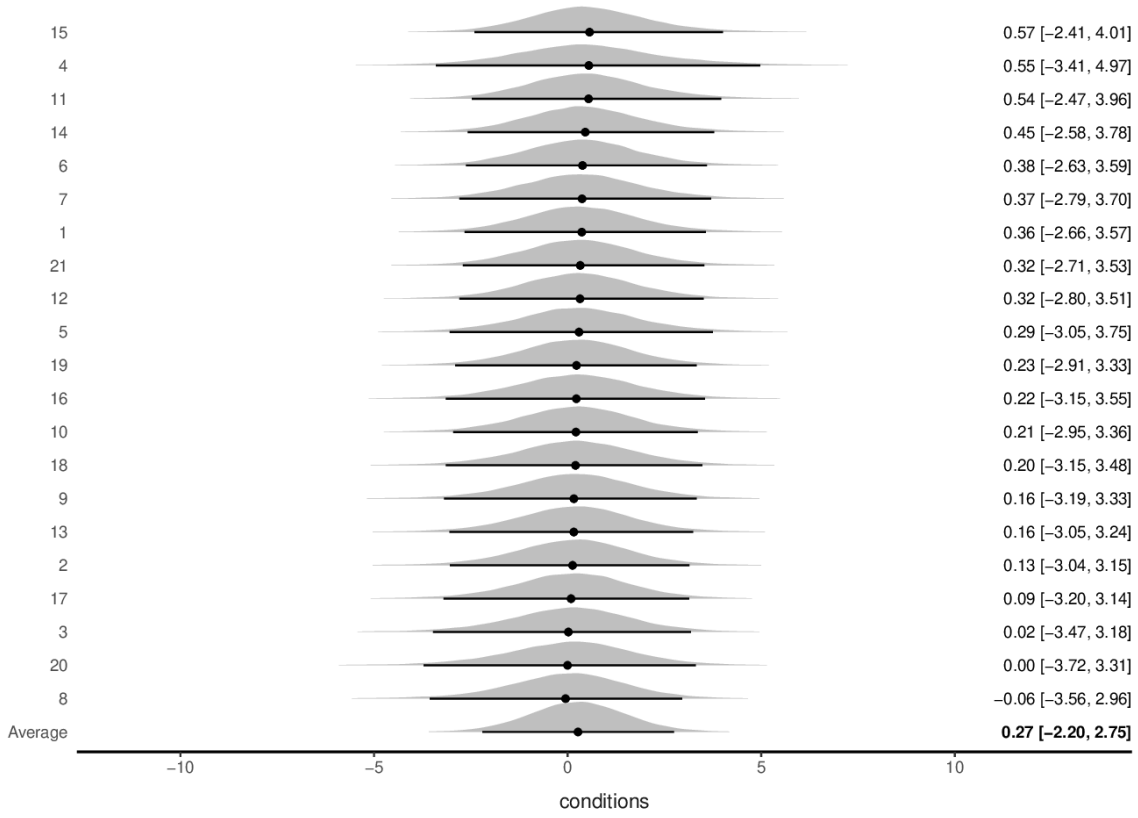


Figure X. shows the group-level effects of the sound condition across participants for the quality score...

Report



[1] http://www.themqm.info/

- Write about errors introduced

Justus, Claudia, Miguel

4. Discussion and conclusion

All

The introduction of the sound condition on the translation workflow has a x effect on cognitive load, a x effect on productivity, and a x effect on quality. Moreover, the effect of the PEMT speed on quality under sound conditions is....

The typed of text together with the sound condition has an effect on all of the measurements, a possible explanation is de complexity of the text...

Limitations

In our design we prioritised ecological validity over internal validity. In doing so, we tried not to deviate too much from the normal working conditions of professional translators and therefore enabled them to work on texts of reasonable size inside a familiar CAT tool environment, and to consult as many external resources as they felt necessary.

This meant that our eye tracking data collection was at the level of target and source text areas, and not at segment-level.
We used standard tests to assess the difficulty of our source texts, but we are aware that these are not always conclusive when it comes to translation difficulty. In the future we are planning on also integrating human evaluators for further assessment of the difficulty levels.

References