Translations and Open Science

Study on machine translation evaluation
in the context of scholarly communication

D3: Outcome for discipline

"Neuroscience and Disorders of the Nervous System"

Version: final

Authors:

Tom Vanallemeersch, Sara Szoc, Kristin Migdisi, Laurens Meeus (CrossLang)

Lieve Macken, Arda Tezcan (LT3)

OPERAS
open scholarly communication in the european research area for social sciences and humanities

CROSSLANG

language and translation technology team

DISCLAIMER

The ideas and views expressed in the exploratory reports only reflect those of the experts involved in the studies and may not be representative of the opinions or policies promoted by any specific organization, institution, or government entity. The present report is therefore only intended for informational purposes.

AVERTISSEMENT

Les idées et les perspectives exprimées dans les rapports exploratoires reflètent uniquement celles des spécialistes ayant contribué aux études et ne sont pas nécessairement représentatives des opinions ou des politiques promues par une organisation, une institution ou une entité gouvernementale spécifique. Le présent rapport est donc uniquement diffusé à des fins d'information.

Table of contents

1. Introduction

This deliverable outlines the evaluation outcome and best practices for the discipline “Neuroscience and Disorders of the Nervous System” (ERC code LS05). In particular, it provides the following description of the data, models and results obtained for this discipline using the procedure outlined in deliverable D1: statistics regarding the training and test material selected for this discipline, information concerning the engines trained, and scores produced using automated MT metrics.

This document is structured in the same way as D1. In Sections 2 to 4, we provide a summary of the information on training and fine-tuning engines, automatic evaluation, and human evaluation, for the discipline in question. Section 5 provides conclusions, while the annexes provide detailed information.

2. Training and fine-tuning MT engines

2.1. Training and evaluation data

The data selected in call 1 consists of seven publication types from 251 different sources of publication, and a terminology list. Table 1 gives an overview of the size and distribution.

Type of publication	Documents	Segments
Article	170	8085
Conference paper abstract	31	159
Journal article abstract	1395	7893
Report	8	938
Research journal article	62	1778
Review abstract	947	23575
Thesis abstract	4860	60747
Terminology	-	415
Total	**7473**	**103591**

Table 1 - Dataset statistics (data from call 1)

Given the preference for texts with an open license (see deliverable D1), the evaluation data is composed of the texts having a CC BY license (e.g. CC BY-SA-4.0) as well as 816 additional abstracts obtained from the **ANR dataset**[1] falling under discipline LS05 (*Neuroscience and Disorders of the Nervous System*). Moreover, we obtained **additional links** to bilingual publications (3 reviews and 3 papers) from OPERAS.

To accommodate non-specialists conducting the self-paced reading experiments, the texts coming from popularised science publications were selected based on the readability level.

2.2. Data partitioning

The dataset for LS05 from call 1 as outlined above was split into training, validation, testing and evaluation sets according to the principles described in Section 3 of deliverable D1. Figure 1 shows the total number of segments used for each subset.

[1] This data was collected from ANR (Agence Nationale de la Recherche, https://anr.fr). See D1 for more details.

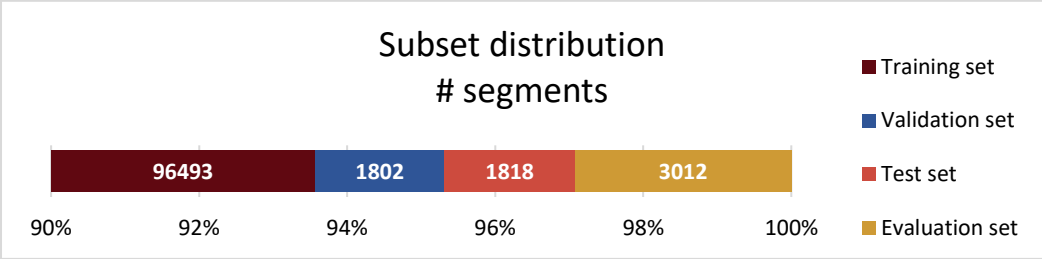

Figure 1 - Distribution of training, validation, testing, and evaluation sets

Regarding the composition of the subsets, the following comments should be made:

Training set: It consists entirely of data from call 1. The aim is to keep as much data as possible in this dataset, while being able to draw statistically significant conclusions for the other subsets.

Validation set: It consists entirely of data from call 1. In order not to split up documents while still having a fair representation of the different types, we made sure to include thesis abstracts, articles, review abstracts and journal article abstracts. For each type, we used 400-500 segments, resulting in around 1900 segments for validation.

Test set: The same criteria as for the validation set apply.

Evaluation set: See Section 2.1.

The composition of the subsets is shown in Figure 2. Annex I provides an overview of the dataset challenges, with examples.

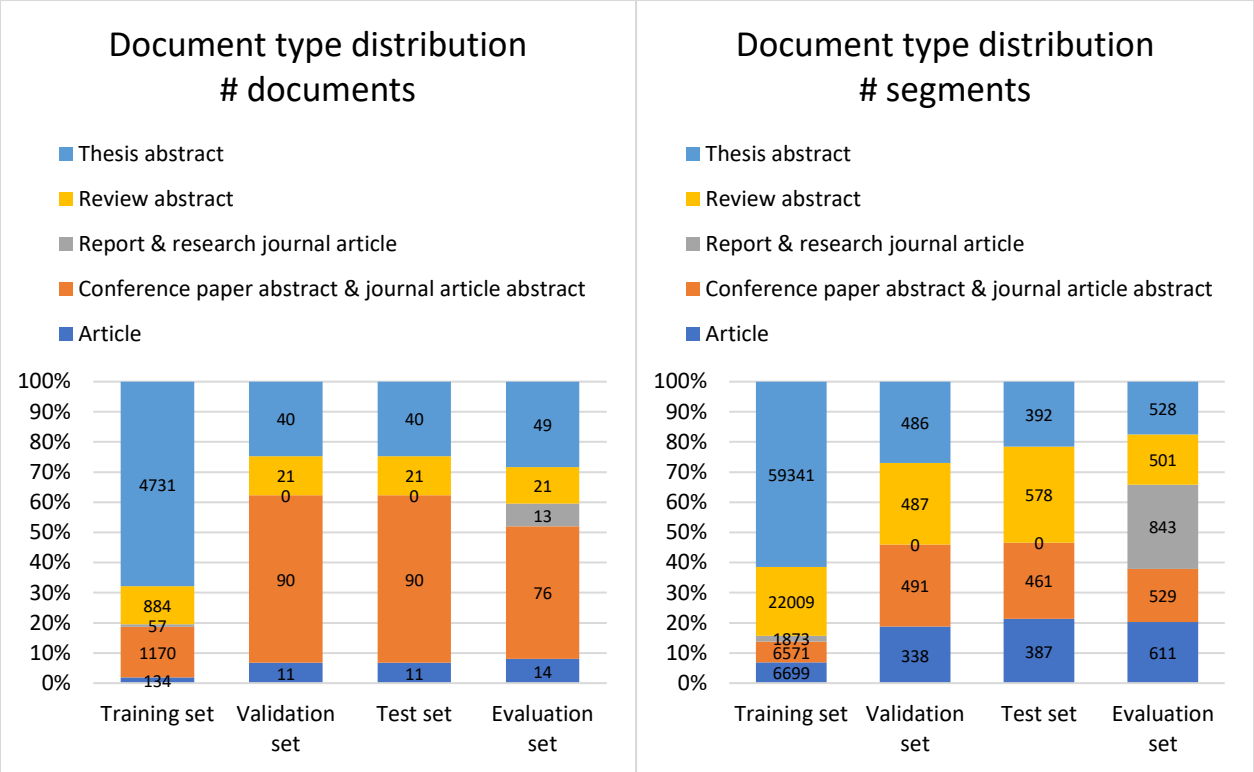

Figure 2 – Distribution of publication types for each subset, number of documents and segments

2.3. MT Customisations

Table 2 gives an overview of the different experiments. Validation set scores for OpenNMT trainings can be found in Annex II. In addition, we translated test sets using eTranslation (cf. Section 3).

Type	System	Short description[2]	Duration[3]	Date
commercial	DeepL	Baseline	/	19/04/2023
		custom (termbase)	5 seconds	19/04/2023
	ModernMT	Baseline	/	19/04/2023
		custom (OPERAS training data)	1 minute 30 seconds	19/04/2023
open source	OpenNMT	Baseline	3 h 20 m/iteration	21/04/2023
		custom 1 (OPERAS training data)	3 h 20 m/iteration	21/04/2023
		custom 2 (OPERAS training data + SciPar)	3 h 20 m/iteration	21/04/2023

Table 2 - Overview of the MT experiments

[2] Baseline refers to the off-the-shelf MT engines (for DeepL and ModernMT) or the MT model trained without any domain-specific training data (for OpenNMT). OPERAS means the engine was trained with the data described in Section 2. SciPar means that the OPUS SciPar dataset (consisting of around 9M segments from scientific abstracts in various domains) mentioned in deliverable D1 was used as additional data to train the engine.

[3] This column gives an idea of the time needed to "fine-tune" (in case of DeepL and ModernMT) or "train" (OpenNMT) the models. For OpenNMT, all trainings were performed on a single NVIDIA GeForce RTX 2080 Ti GPU with 11 GB of memory.

3. Automated evaluation

Each MT system was scored using a set of automatic metrics, as described in Section 3 of deliverable D1. One of these metrics is BLEU (the SacreBLEU variant), the results for which are shown in Figure 3. It indicates that there is hardly any difference between the DeepL baseline and DeepL using the termbase. The disparity is slightly larger for ModernMT baseline versus fine-tuned, while OpenNMT shows a much more pronounced difference between baseline and fine-tuned, with the engine making use of SciPar in its training data generally performing the best. DeepL performs best for articles and journal article abstracts, while OpenNMT appears with SciPar performs best in case of review abstracts and thesis abstracts. Finally, eTranslation scores are slightly lower compared to OpenNMT fine-tuned without SciPar data in case of articles and journal article abstracts.

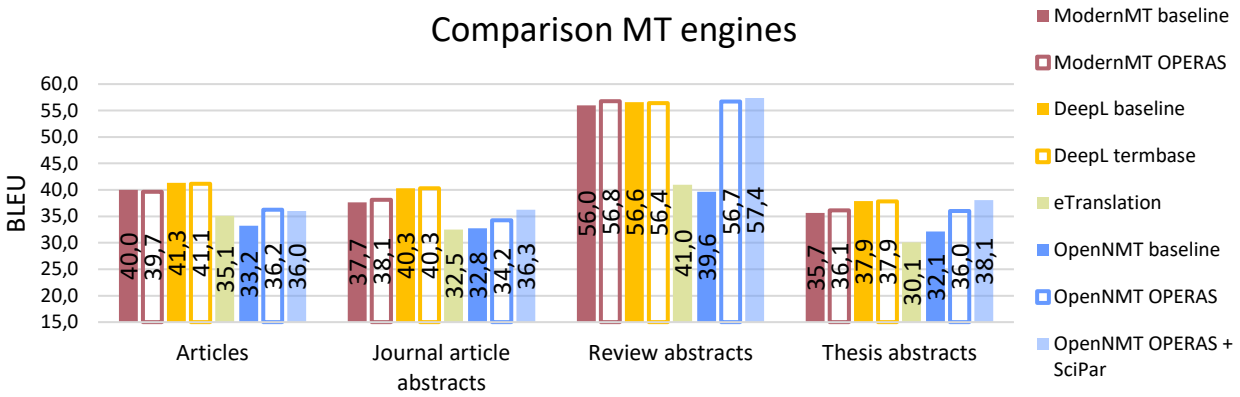

Figure 3 – Comparison of MT engines, using BLEU score, for each text type

Similar observations are made when applying other metrics (TER, ChrF, METEOR and COMET). These results are shown in Annex II:

- The TER, METEOR and ChrF scores are generally in line with the ones from BLEU: when an engine has a higher BLEU score than the baseline, it also tends to have a lower TER score and a higher METEOR score.
- The picture for COMET scores is more variable.
- The scores hardly change between the first 30 epochs and the 60th epoch. This is also the case for the validation set.

Based on the above observations for various metrics, we decided to perform human evaluation for 3 engines: the DeepL baseline, the fine-tuned ModernMT engine, and the OpenNMT engine fine-tuned with in-domain data and the SciPar dataset.

4. Human evaluation

After setting up paragraph samples based on the procedure described in Section 4.2 of deliverable D1 and the evaluation set described above (Section 2.2), we set up the tasks, contacted the evaluators, followed up on the execution of the tasks, and processed and interpreted the results.

4.1. Setup and execution of adequacy task

MT-Eval batch files were set up following the procedure outlined in Section 4.3 of deliverable D1: sampling of appropriate paragraphs, listing them in random order, translating them using the three selected engines mentioned in Section 3 above, manually checking the source segments, MT outputs and reference translations, and converting the source segments and the MT output to MT-Eval batch files.

The evaluations were performed by two professional translators and two researchers working at the University of Aix-Marseille. More details about the evaluators and the feedback received can be found in Annex III.

4.2. Results of adequacy task

Based on the evaluation outcome (enriched CSV files), we produced a number of statistics. For a comprehensive understanding of the adequacy task, please refer to Annex III, which contains a detailed overview. In the present section, we present a concise summary of the results.

User ratings

When looking at the user ratings, we conclude with significant confidence that DeepL is on average higher rated than ModernMT, which is in turn higher rated than OpenNMT. We also notice that researchers rate the translations on average higher than the translators. We cannot say with significant confidence that the average rating differs between the document types.

Number of times each engine is ranked first

Another statistic we produced relates to the MT engine rankings implictly assigned by evaluators through the ratings they provided. The results show that DeepL clearly performs best in this perspective, as it is ranked much more often as sole best system than the other two engines, and is also involved in many ties.

Correlations

When investigating the correlation between automatic metrics and human ratings, we notice there is a low correlation between the BLEU score and human ratings. Nevertheless, a higher BLEU score tends to lead to a higher human rating.

4.3. Post-editing task

Based on the evaluation outcome (enriched CSV files), we produced a number of statistics. These are available in Annex IV. Below, we present a summary of the most interesting findings.

Post-editing times

When examining the post-editing times, we observe a large range, from a couple of seconds to tens or even hundreds of seconds for each evaluator. We notice that the translators take on average much longer to correct the text than the researchers. One possible explanation for this could be that the translators are more strict when it comes to correcting the translation.

The post-editing times per engine show that DeepL produces better outputs than ModernMT, and the latter, in turn, produces better outputs than OpenNMT. However, it appears that the post-edit times do not strongly differ between MT engines.

When we look at the post-editing times per document type, we see that abstracts took on average the longest to edit.

Perceived effort

When we look at the MT engines in terms of perceived effort, we can say with confidence that post-editing DeepL outputs has a lower average perceived effort than ModernMT outputs, which in turn has a lower average effort than OpenNMT outputs. This is in correspondence to the ranking of engines based on the automatic evaluation results.

The comparison of perceived efforts confirms the previous findings. Abstracts have a higher perceived effort than articles.

When comparing post-editing time and perceived effort, we can say with significant confidence that there is a correlation between them. Even though evaluators had a large difference in average post-editing time, the perceived effort still correlates well with post-editing time. We cannot say with significant confidence that the median post-editing times for a perceived effort of 4 and 5 differ.

HTER

When calculating the HTER and comparing it with the perceived effort, we can clearly see a correlation. While the median HTER of a perceived effort of 5 appears to be lower than for a perceived effort of 4, we have too few samples to make any significant conclusions for this.

Finally, we can see that there is a correlation between post-editing time and HTER.

4.4. Self-paced reading experiment

Twelve texts were selected for the discipline from three different sources (see Table 3). It was rather difficult to select suitable texts for lay persons as the discipline contained rather technical texts. Moreover, as the texts had to be similar in length, coherent text excerpts were occasionally manipulated by deleting intermediate sentences. No abstracts were available for the discipline. We therefore created two sets of full articles (one with only texts selected from a popularised science publication, and one with a mixture of articles coming from different sources). The two sets of full articles were analyzed together.

NEUROSCIENCE	No. src words	No. segments
COCHRANE SYSTEMATIC REVIEWS		
6_article	136	7
4_article	139	6
OPERAS_000093_LS05_RA	127	5
OPERAS_000841_LS05_RA	144	5
FULL ARTICLES (THE CONVERSATION)		
OPERAS_003898_LS05_ART	192	7
OPERAS_003926_LS05_ART	201	8
OPERAS_003943_LS05_ART	197	8
OPERAS_003943_LS05_ART	187	9
FULL ARTICLES		
OPERAS_003947_LS05_ART	129	6
Google_doc/1_article	128	7
Google_doc/4_article	124	6
OPERAS_003947_LS05_ART	136	6
TOTAL	**1840**	**64**

Table 3 - Data selection for the self-paced reading experiment

Twelve UGent staff members (within the age range of 23-51 years old) participated in the experiments. All participants are highly proficient in French and are used to reading academic articles. All participants signed an informed consent form and got a financial reward of 20€ .

The experiments were carried out from June 1st until June 19th, 2023 (during the same time span, the experiments for discipline 3, see D4, were also carried out). The duration of the sessions varied between 60 and 80 minutes.

Translation quality was assessed as sufficient in 81% of all assessments (which is higher than discipline 1, i.e. Human Mobility, where the result was 74%). In 27 of the 144 assessments, translation quality was rated as insufficient, as indicated in Table 4.

Quality Score: no		Total
	DeepL	4
	ModernMT	5
	HT	6
	OpenNMT	12
Total		**27**

Table 4 - Number of translations rated as insufficient, per engine

Average normalized reading times (ms per word) were lowest for ModernMT (454 ms) and DeepL (474 ms), higher for by HT (481 ms), and highest for OpenNMT (546 ms), although there is some variation across text types (see Figure 4), with longer reading times for the Cochrane reviews. These results are comparable to discipline 1.

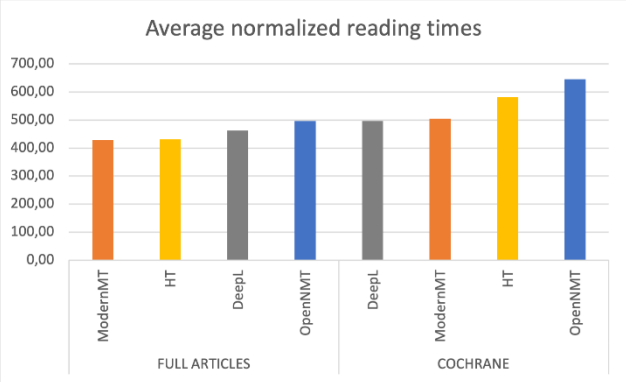

Figure 4 - Average normalized reading times

4.5. MQM error annotation

The same dataset that has been used for the self-paced reading experiments was manually analyzed for machine translation errors. Prior to error annotation, terms were marked in the source texts (a) using the term lists provided per domain, and (b) by the annotator. The number of terms marked during both steps are as follows:

- Terms marked using the term list LS5_Neuroscience.tsv: 0
- Terms marked by the annotator: 120

After the MQM error annotation was made on Label Studio, the results were analysed per text type and for the whole evaluation set. These results are presented in two categories: (i) MQM scorecards, and (ii) other analyses.

The MQM scorecards regarding all evaluation data, per MT engine, are provided in Annex V. We also provide the scorecards per text type, per engine (.xlsx) in a separate zip file. The results of other analyses are provided per text type and for the whole evaluation set, per MT engine, in Figure 5. The information in the graph with MQM scores and in the graph with ratio of sentences with errors is also in present in the MQM scorecards.

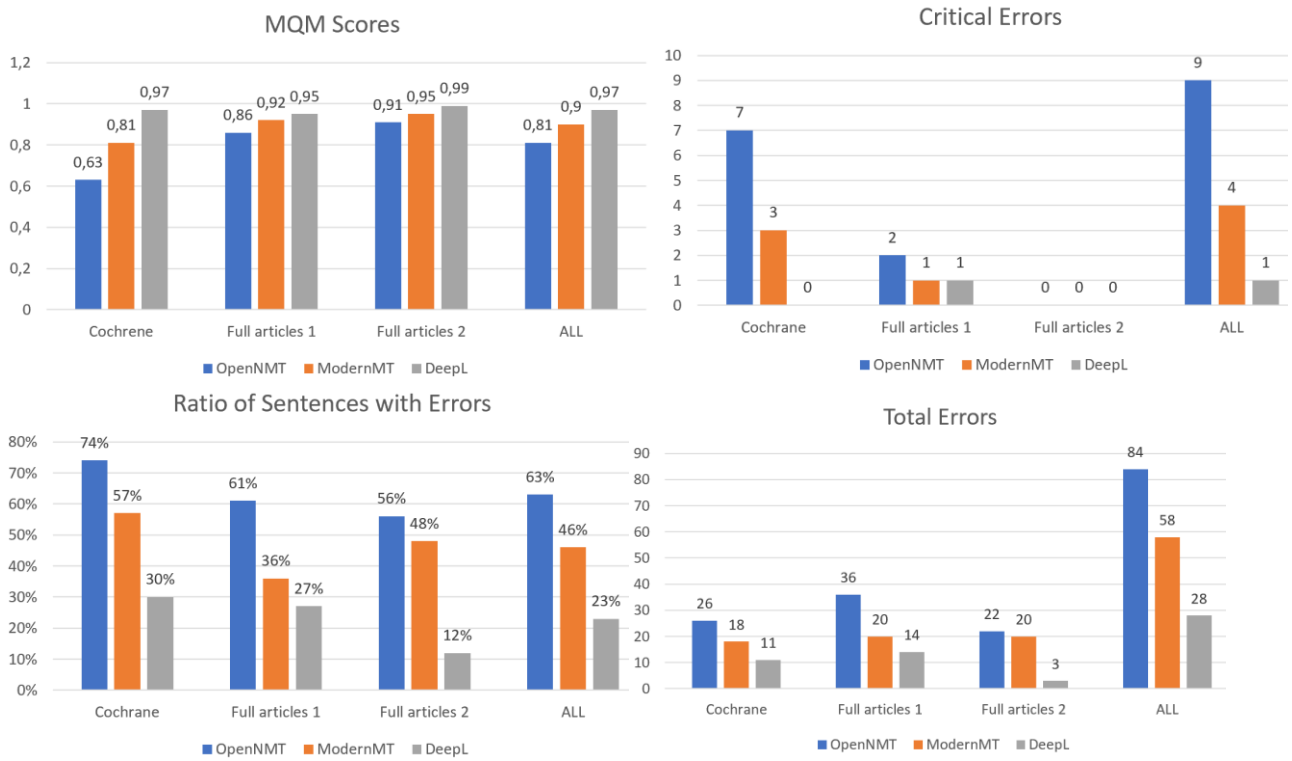

Figure 5 – Various types of scores resulting from manual error annotation

From the scorecards and analyses, we can conclude that we obtain the same ranking of engines as in case of automatic evaluation scores, i.e. DeepL scores better than ModernMT and ModernMT scores better than OpenNMT. We also observe differences in scores per document type. For instance, Cochrane translations have a clearly higher ratio of sentences with errors than other document types.

5. Conclusions

In this deliverable, we presented detailed information on the second discipline “Neuroscience and Disorders of the Nervous System”, more particularly regarding the data, models and results obtained. Using domain-specific data, we customised both open-source (OpenNMT) and commercial MT systems (DeepL and ModernMT) and partitioned the data into training sets, evaluation sets, test sets and validation sets.

Each MT system (as well as the eTranslation system) was scored using a set of automatic metrics. The automatic scores showed no clear difference between DeepL baseline and DeepL using a termbase. This difference was slightly larger for ModernMT baseline and fine-tuned. The most significant difference was observed for OpenNMT fine-tuned (with and without SciPar data) and baseline. Overall, the scores for DeepL were the highest. In addition to the automatic scores, human evaluations were performed. Four types of tasks were performed in order to obtain the results (adequacy task, productivity task, self-paced reading experiment and MQM error annotation).

The adequacy task showed the highest rating for DeepL, followed by ModernMT and OpenNMT. DeepL is also more often ranked as sole best system. Moreover, a low correlation is seen between the BLEU score and human ratings. Nevertheless, a higher BLEU score tends to lead to a higher human rating.

Results from the productivity task indicate that DeepL produces the best outputs. However, in terms of post-editing time, there is no significant difference between the engines. Translators take on average much longer to correct the text than the researchers. Abstracts took on average the longest to edit. Furthermore, post-editing DeepL outputs showed the lowest average perceived effort, followed by ModernMT and OpenNMT. A correlation was observed between perceived effort and post-editing time, between perceived effort and HTER, and between HTER and post-editing time.

Regarding the self-paced reading experiment, it was rather difficult to select suitable texts for lay persons as the discipline contained rather technical texts. Translation quality was assessed as sufficient in 81% of all assessments. Average normalized reading times (milliseconds per word) were lowest for ModernMT (454 ms) and DeepL (474 ms), higher for by HT (481 ms), and highest for OpenNMT (546 ms).

From the MQM scorecards and analyses, we can conclude that we obtain the same ranking of engines as in case of automatic evaluation scores, i.e. DeepL scores better than ModernMT and ModernMT scores better than OpenNMT. Differences in scores per document type were also observed.

Annex I: Dataset challenges and examples

This annex gives an overview of the challenges encountered when working with the provided datasets throughout the various phases of the project: understanding the data, dataset preprocessing, model training, setting up automatic and human evaluation, and results processing. We present a breakdown of the various issues that arose, accompanied by relevant examples to illustrate these challenges.

Segmentation issues

Fragments of journal articles with sentences glued together:

Source EN	Reference FR
They are sometimes unpredictable and sometimes on **cue.They** can surprise us, stimulate us, move us to action and sometimes to tears.	Elles peuvent nous surprendre, nous stimuler, nous pousser à l'action et parfois aux larmes.

Source EN	Reference FR
These neural networks are ordinarily inactive, but when they are excited by other brain activity, such as a stimulus, a related thought or hunger, they compete for access to consciousness based on their **strength.The** competitive strength of networks is influenced by their relevance to our situation, but also to our goals, needs, interests or emotions.	Les réseaux neuronaux stimulés se font concurrence pour accéder à la conscience et la force concurrentielle des réseaux est influencée par leur pertinence par rapport à notre situation, nos objectifs, nos besoins, nos intérêts ou nos émotions.

Source EN	Reference FR
Passion induces positive spontaneous **thoughts.Even** during uneventful daily activities, weak emotions or microemotions such as worries, desires, irritation, stress, surprise or interest are involved in orienting many of our thoughts .	Même quand nous ne vivons pas d'émotions fortes, il arrive que de faibles émotions, ou microémotions, telles que les inquiétudes, les désirs, l'irritation, le stress, la surprise ou l'intérêt activent nos pensées spontanées .

Table 5 - Bad source examples in the data, sentences glued together

Freely translated outputs

Fragment of journal article:

Source EN	Reference FR
Microemotions of guilt or pride trigger moral intuitions of anticipated disapproval or approval of others, which are essential to develop pro-social behaviour such as co-operation, helpfulness and other types of behaviour that benefit others.	Les microémotions de culpabilité ou de fierté déclenchent des intuitions morales de désapprobation ou d'approbation anticipée des autres, qui sont essentielles pour développer un comportement prosocial.

Fragment of journal article abstract:

Source EN	Reference FR
The main contributions of Pankow are the following : the use of non-traditional techniques (face to face, drawings, models, family contacts) in the treatment of the psychotic, the redefinition of the concept of forclusion (Lacan) as a defense mechanism directly implying the body image.	Les principaux apports de Pankow sont les suivants : l'utilisation de techniques non traditionnelles (face à face, dessins, modelages, contacts avec la famille) dans le traitement du psychotique, la redéfinition de l'image du corps avec une double fonction de forme et de contenu, le développement de la technique des greffes de transfert, par opposition au transfert classique, la redéfinition du concept de forclusion (Lacan) en tant que mécanisme de défense impliquant directement l'image du corps.

Table 6 - Example of freely translated outputs

Annex II: Automatic scores

Table 7 provides metric scores for all document types. Table 8 provides validation scores.

		Scores			
Type	**Engine**	**BLEU**	**TER**	**ChrF**	**COMET**
Article	ModernMT baseline	39,99	49,65	66,19	86,70
	ModernMT OPERAS	39,66	50,23	65,92	86,61
	Deepl baseline	41,31	49,02	67,46	87,74
	Deepl termbase	41,13	49,22	67,38	87,74
	eTranslation	35,12	55,31	63,45	82,88
	OpenNMT baseline (30 epochs)	33,21	56,08	61,96	83,12
	OpenNMT OPERAS (30 epochs)	36,23	53,40	62,99	83,34
	OpenNMT OPERAS + SciPar (30 epochs)	36,01	53,77	63,20	83,66
	OpenNMT baseline (60 epochs)	34,77	55,01	62,86	83,21
	OpenNMT OPERAS (60 epochs)	34,86	54,53	62,44	83,8
	OpenNMT OPERAS + SciPar (60 epochs)	34,65	54,68	62,53	83,78
Journal article abstract	ModernMT baseline	37,67	52,28	67,43	87,31
	ModernMT OPERAS	38,14	51,52	67,69	87,35
	Deepl baseline	40,32	49,51	68,92	88,27
	Deepl termbase	40,31	49,50	68,93	88,26
	eTranslation	32,49	57,52	64,53	85,60
	OpenNMT baseline (30 epochs)	32,76	56,90	64,23	84,76
	OpenNMT OPERAS (30 epochs)	34,24	55,06	65,10	85,35
	OpenNMT OPERAS + SciPar (30 epochs)	36,25	53,57	65,86	86,18
	OpenNMT baseline (60 epochs)	33,37	56,36	64,43	84,87
	OpenNMT OPERAS (60 epochs)	34,28	55,38	64,80	85,21
	OpenNMT OPERAS + SciPar (60 epochs)	36,35	53,63	65,80	86,29
Research abstract	ModernMT baseline	56,01	36,28	74,72	86,30
	ModernMT OPERAS	56,77	36,64	75,46	86,68
	Deepl baseline	56,56	36,66	74,85	87,68
	Deepl termbase	56,42	36,71	74,79	87,68
	eTranslation	41,00	48,75	66,51	83,84
	OpenNMT baseline (30 epochs)	39,64	50,12	65,67	82,78
	OpenNMT OPERAS (30 epochs)	56,71	35,37	76,73	87,21
	OpenNMT OPERAS + SciPar (30 epochs)	57,35	35,00	76,60	87,25
	OpenNMT baseline (60 epochs)	39,12	50,49	65,54	83,23
	OpenNMT OPERAS (60 epochs)	55,66	36,50	75,95	86,93
	OpenNMT OPERAS + SciPar (60 epochs)	56,13	35,78	76,03	87,24
Thesis abstract	ModernMT baseline	35,66	56,36	65,81	83,09
	ModernMT OPERAS	36,10	55,88	66,02	83,25
	Deepl baseline	37,86	54,95	67,03	83,94
	Deepl termbase	37,85	55,13	67,07	83,90
	eTranslation	30,14	61,84	62,48	81,50

	OpenNMT baseline (30 epochs)	32,11	59,96	63,62	81,07
	OpenNMT OPERAS (30 epochs)	36,00	55,34	65,41	82,20
	OpenNMT OPERAS + SciPar (30 epochs)	38,09	53,67	66,40	83,05
	OpenNMT baseline (60 epochs)	32,14	59,97	63,74	81,58
	OpenNMT OPERAS (60 epochs)	35,71	56,19	64,92	82,20
	OpenNMT OPERAS + SciPar (60 epochs)	37,90	54,05	66,09	82,99

Table 7 - Automatic scores for document types

Validation set **OpenNMT** **SacreBLEU**

Engine	10 epochs	20 epochs	30 epochs	40 epochs	50 epochs	60 epochs
OpenNMT baseline	35.3	36.0	35.8	36.2	36.1	36.1
OpenNMT OPERAS	41.8	42.4	41.6	41.7	41.8	42.0
OpenNMT OPERAS + SciPar	43.2	43.4	43.2	43.1	42.9	42.9

Table 8 - BLEU score on validation set for every 10 iterations

Annex III: Adequacy task

Setup and execution

The contact details of two professional translators were provided by the University of Rennes. OPERAS provided the contact details of two researchers working at the University of Aix-Marseille (one native speaker of French and one native speaker of Serbian). We decided to reduce the envisaged number of segments from the planned 500 per task to 400 for time and budget reasons for the translators and the number of segments for the researchers to 200, and proposed a price to the evaluators and a time span of two weeks for performing the work. Depending on the evaluator, the price for the adequacy task was based on an estimate of 1 minute per segment and an hourly rate (the work amounting to more or less 7 hours) or fixed. After the people contacted agreed with the conditions, we provided them with the instructions for performing the task, the MT-Eval links, a bilingual terminology list, abstracts relating to the segments to be evaluated, CrossLang's standard NDA to sign, and, in case of the researchers, a service contract to sign.

Some of the evaluators provided feedback relating to the tasks:

- One translator indicated that the tasks should be less complex "in real life" as people will be working on whole texts and not bits of them, and that it is very hard to guarantee term consistency when you cannot see the whole picture and go back to previously translated segments.
- One researcher wondered how the translation of source segment should be evaluated when the latter is bad (cfr. a remark of a translator in D2).

We followed up on the progress of the evaluator's work directly in MT-Eval, as the tool keeps track of the number of segments evaluated. All evaluators performed their work in the time frame agreed upon.

Detailed results

The graphs in Figure 6 show the distribution of all evaluators' ratings (ranging from 1 to 5, i.e. very poor to excellent) and the distribution for each type of evaluators separately, i.e. translators (1, 2) and researchers (3, 4). From the user ratings, we can conclude with significant confidence that DeepL is on average higher rated than ModernMT, which is in turn higher rated than OpenNMT. We also notice that researchers rate the translations on average higher than the translators.

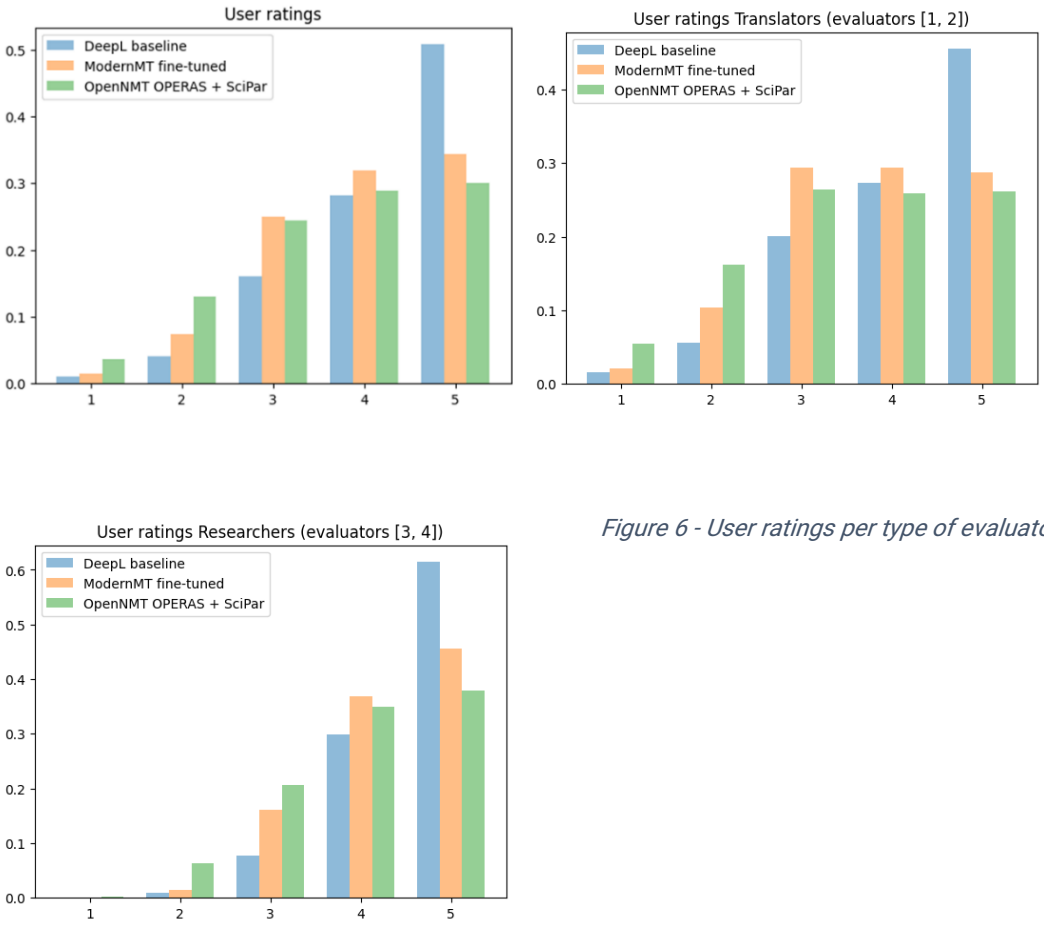

Figure 6 - User ratings per type of evaluator

Figure 7 shows the distribution of all evaluators' ratings per document type. We cannot say with significant confidence that the average rating differs between the types.

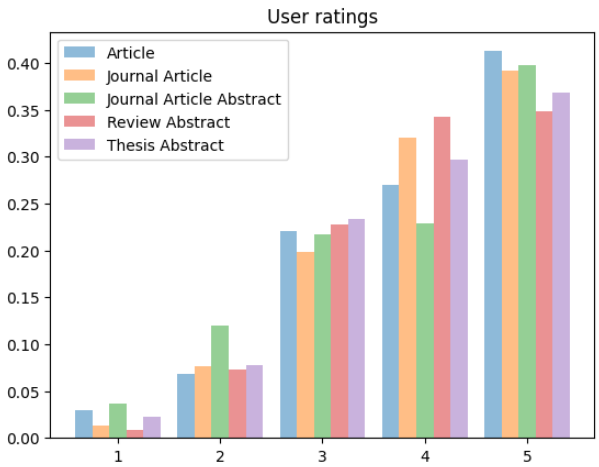

Figure 7 - User ratings per document type

Another statistic we produced relates to the MT engine rankings implicitly assigned by evaluators through the ratings they provided. This is shown in Figure 8, which presents the number of times a specific engine was ranked first for a given segment. The bright, bottom part depicts the number of times it was ranked better than both other engines, while the darker, top part depicts the number of times there was a tie between two or more engines. The DeepL engine clearly performs best in this perspective, as it ranked much more as sole best system than the other two engines, and is also involved in many ties.

When investigating the correlation between automatic metrics and human ratings, shown in the graphs in Figure 9, we notice there is a low correlation between BLEU score and human ratings. Nevertheless, a higher BLEU score tends to lead to a higher human rating.

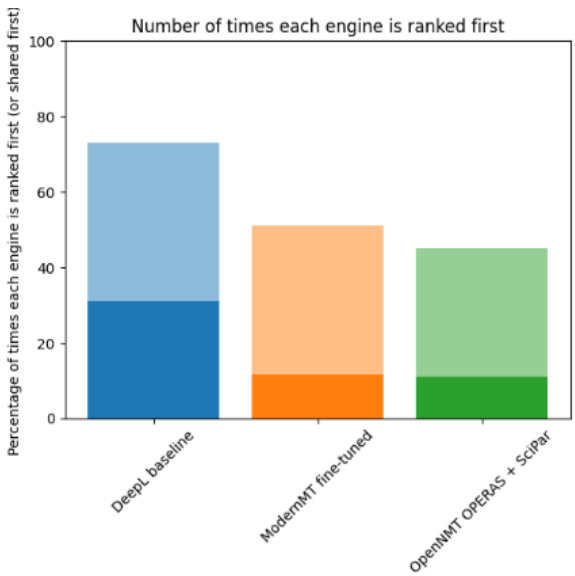

Figure 8 - Number of times engines are ranked first

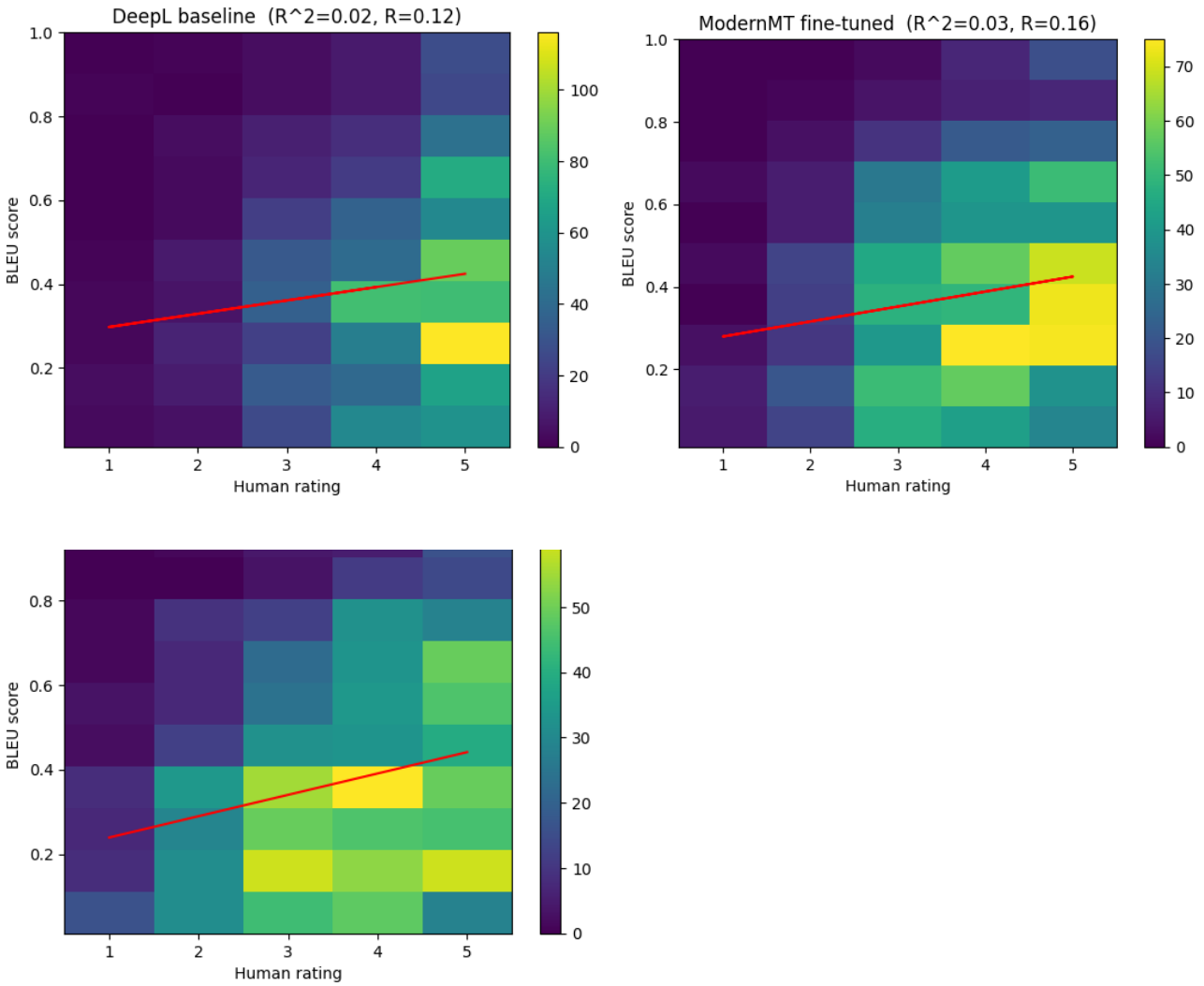

Figure 9 - Correlation between automatic metrics and human ratings

Annex IV: Productivity task

Setup and execution

MT-Eval batch files were set up following the procedure outlined in Section 4.4 of deliverable D1.

The task was performed by the same two professional translators and the same two researchers as those executing the adequacy task. We decided to reduce the envisaged number of segments from the planned 500 per task to 400 for time and budget reasons in case of the translators and to 200 in case of the researchers, and proposed a price to the evaluators and a time span of two weeks for performing the work. Depending on the evaluator, the payment was per hour or fixed. The number of hours (15) required for post-editing was estimated using the average sentence length of the segments involved and a post-editing speed of 750 words per hour (after consultation with University of Rennes). After the people contacted agreed with the conditions, we provided them with the instructions for performing the task, the MT-Eval links, a bilingual terminology list, abstracts relating to the segments to be evaluated, CrossLang's standard NDA to sign, and, in case of the researchers, a service contract to sign.

Detailed results

Figure 10 shows the distribution of the post-edit time for each of the evaluators, i.e. translators (1, 2) and researchers (3, 4). The median post-edit time is provided, together with a confidence interval of the median. Each evaluator has a large range of post-editing times, from a couple of seconds to tens or even hundreds of seconds.

Due to the large range of post-edit times, we worked in the logarithmic domain for all the following calculations.

Y = log10(X), with X being the post-edit time

SEM_Y = SEM(Y)

Confidence interval log10 = [Y_MEDIAN – SEM_Y, Y_MEDIAN + SEM_Y]

Confidence interval = [10**(Y_MEDIAN – SEM_Y), 10**(Y_MEDIAN + SEM_Y)]

One thing we notice is that the translators take on average much longer to correct the text than the researchers. One possible explanation for this is that the translators are more strict when it comes to correcting the translation.

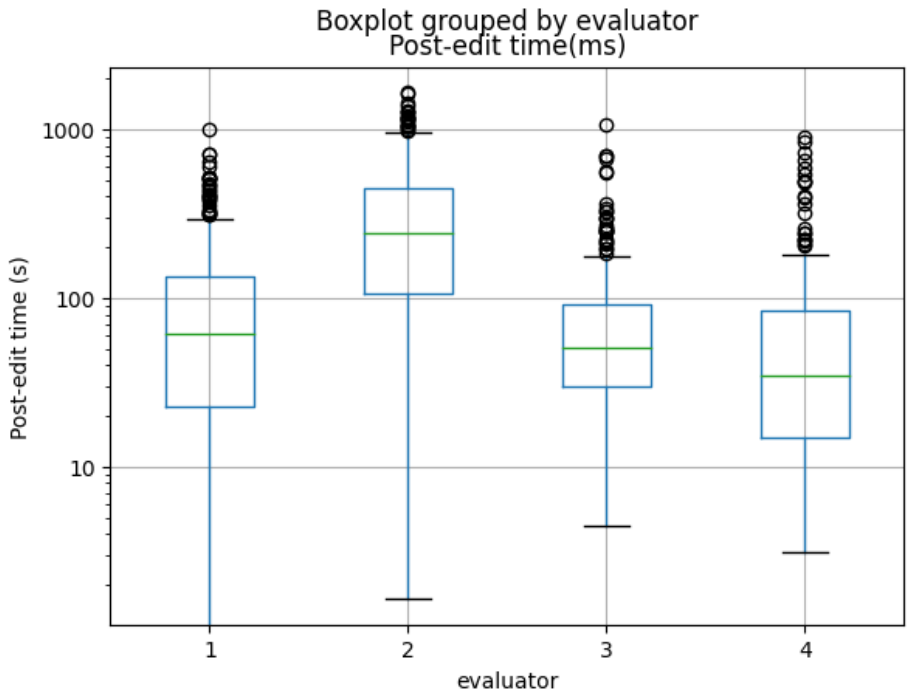

Figure 10 - Boxplot grouped by evaluator - post-edit time (ms)

When investigating the correlation between post-edit time and perceived effort, we obtain Figure 11. It shows the median post-edit time together with a confidence interval of the median. Figure 12 shows the individual evaluators' graphs for clarity. Even though there is still a large range of post-edit times for each group of perceived effort scores, we can say with significant confidence that there is a correlation between perceived effort and post-edit time.

Key takeaways:

- Even though each evaluator had a large difference in average post-edit time, the perceived effort still correlates well with post-edit time.
- We cannot say with significant confidence that the median post-edit times for a perceived effort of 4 and 5 differ. As shown in the individual evaluators' graphs, the perceived effort for 5 is higher or more or less equal than for 4 in case of three evaluators and much lower than for 4 in case of the remaining evaluator.

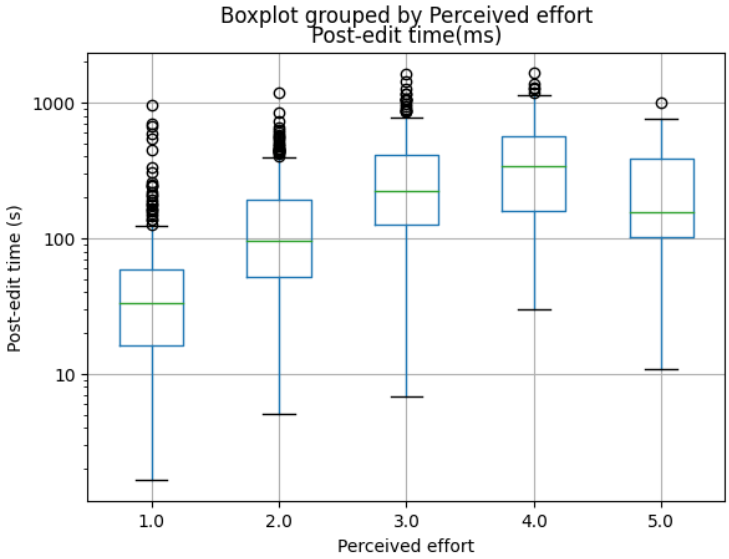

Figure 11 - Boxplot grouped by perceived effort - post-edit time (ms)

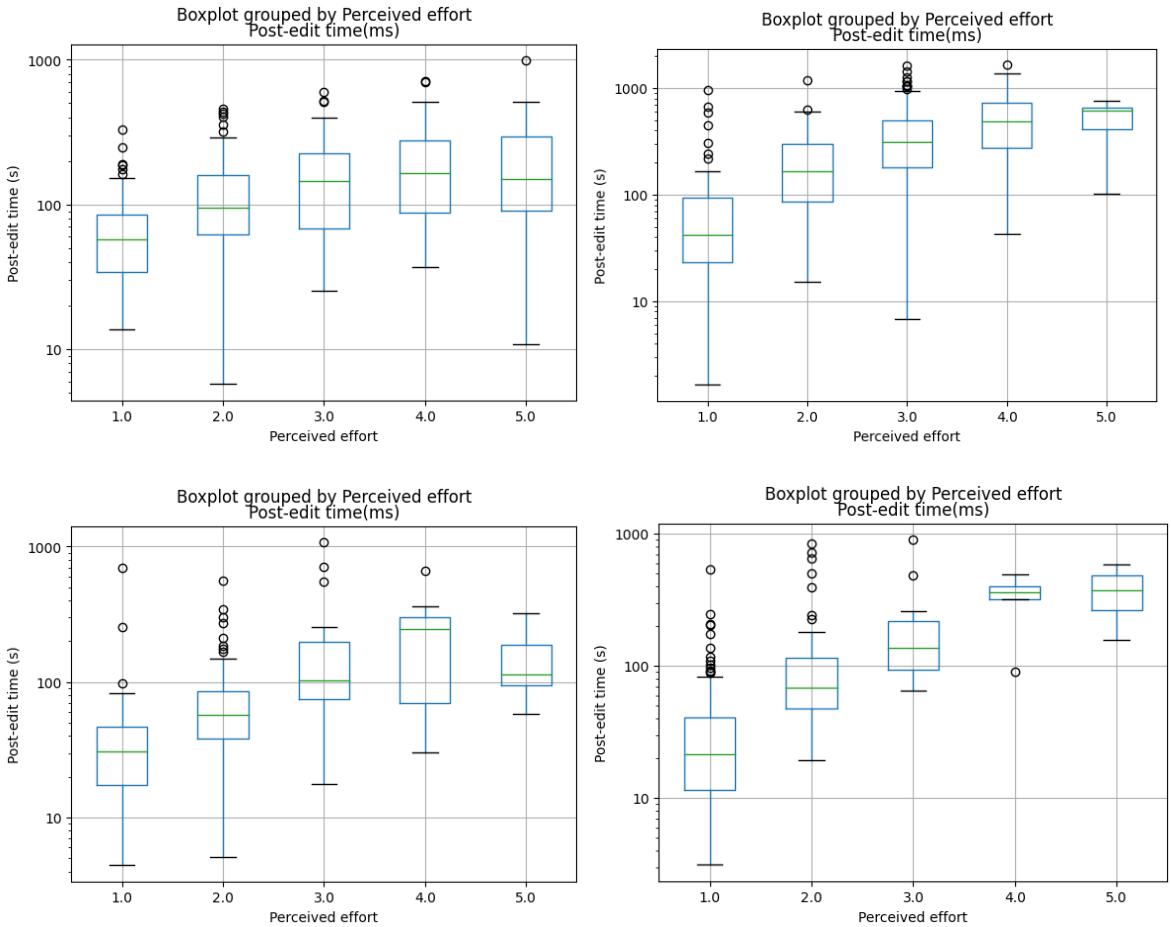

Figure 12 - Boxplot by perceived effort - post-edit time, individual evaluators

Figure 13 shows the post-edit time per engine. From the automatic evaluation we concluded that DeepL produces better outputs than ModernMT, and the latter, in turn, better outputs than OpenNMT. However, it appears that the post-edit times do not strongly differ between MT engines.

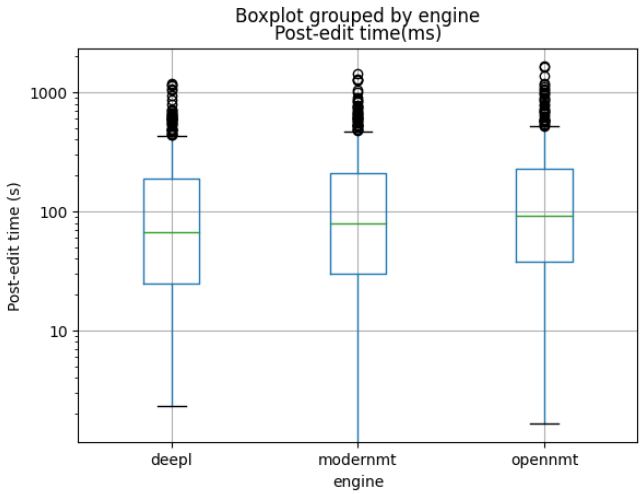

Figure 13 - Boxplot grouped by engine - post-edit time(ms)

In Figure 14, we look at the MT engines in terms of perceived effort. We can say with confidence that post-editing DeepL outputs has a lower average perceived effort than post-editing ModernMT outputs, which in turn has a lower average effort than post-editing OpenNMT outputs. This is in correspondence to the ranking of engines based on the automatic evaluation results.

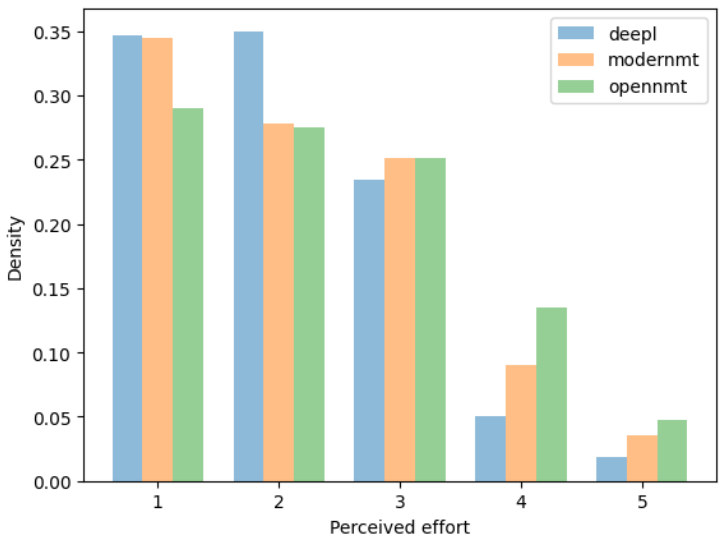

Figure 14 – Perceived effort per engine

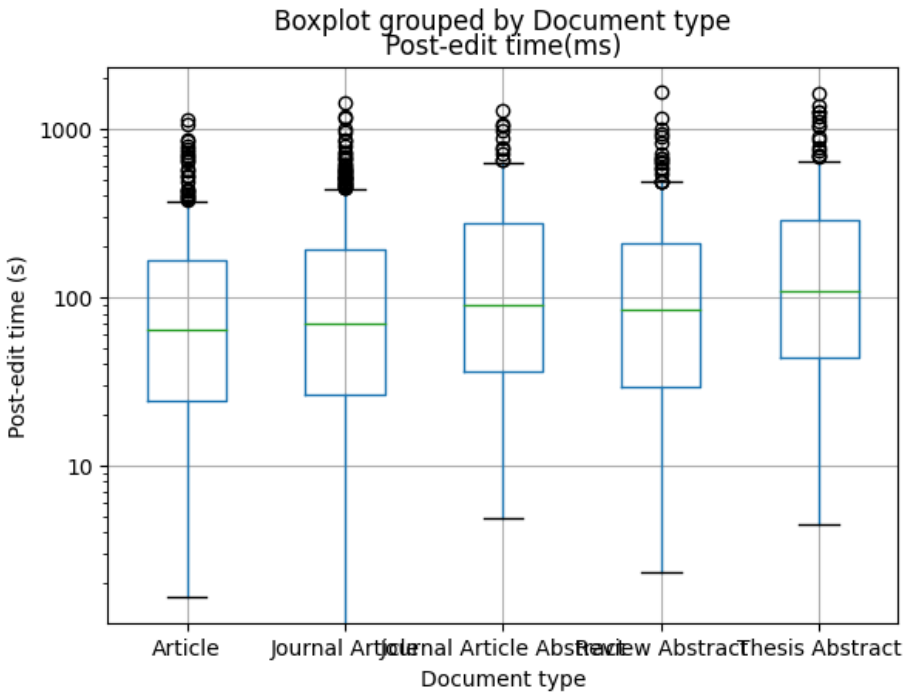

Figure 15 - Boxplot grouped by document type - post-edit time (ms)

Figure 15 shows the post-editing time per document type. Abstracts took on average the longest to edit.

The comparison of perceived efforts in Figure 16 confirms the previous findings. Abstracts have a higher perceived effort than articles.

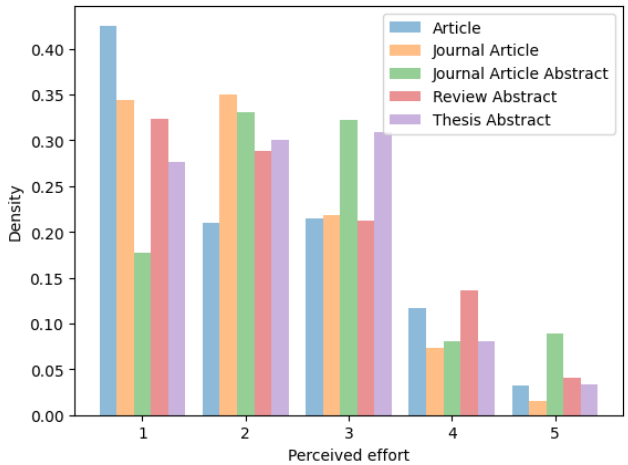

Figure 16 - Perceived effort per document type

When calculating the HTER and comparing it with the perceived effort, we can clearly see a correlation, as shown in Figure 17. While the median HTER of perceived effort 5 appears to be lower than for perceived effort 4 (Figure 18), we have too few samples to make any significant conclusions for this.

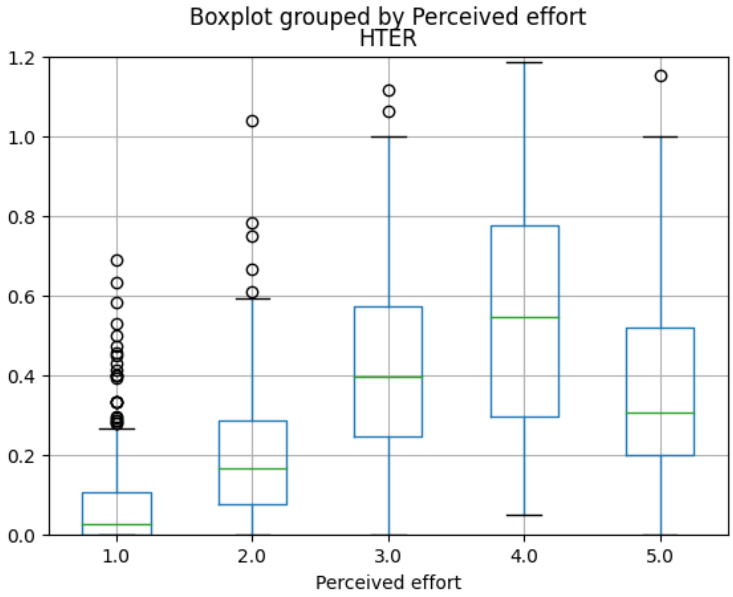

Figure 17 - Boxplot grouped by perceived effort - HTER

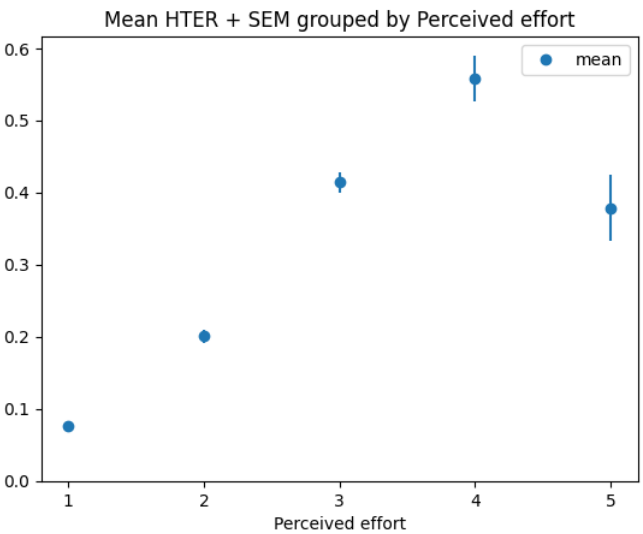

Figure 18 - Mean HTER + SEM grouped by perceived effort

There is a correlation between post-editing time and HTER, as illustrated in Figure 19.

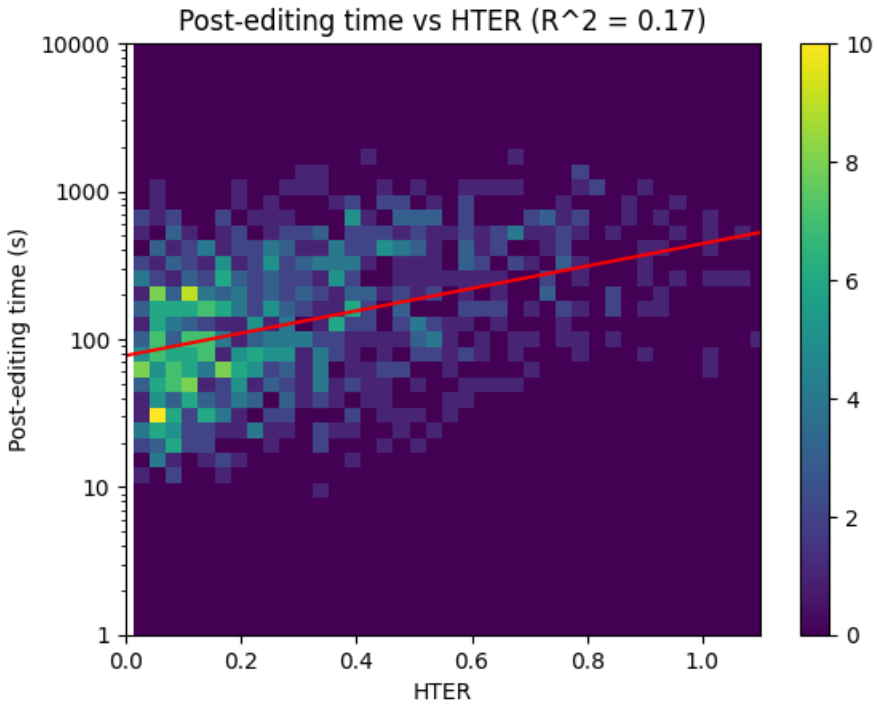

Figure 19 - Post-editing time vs HTER

Annex V: MQM error annotation results

The MQM scorecards regarding all evaluation data, per MT engine, are provided below.

Domain	MT System					
D2 ALL	OpenNMT					
Error Severity Levels:	Neutral	Minor	Major	Critical	Error Type Penalty Total	
Severity Multipliers:	0	1	5	25		
Error Types	Error Counts				ET Weights	ETPTs
Term_Resource	4	7	1	0	1	12.0
Term_Inconsistent	0	1	0	0	1	1.0
Term_Wrong	0	0	1	1	1	30.0
Acc_Mistrans	1	5	15	2	1	130.0
Acc_Overtrans	0	0	0	0	1	0.0
Acc_Undertrans	0	1	0	0	1	1.0
Acc_Add	0	0	0	0	1	0.0
Acc_Omi	0	0	1	3	1	80.0
Acc_DNT	0	0	0	1	1	25.0
Acc_Untrans	0	0	0	0	1	0.0
Ling_Grammar	0	8	5	0	1	33.0
Ling_Punct	0	1	1	0	1	6.0
Ling_Spelling	0	0	0	0	1	0.0
Ling_Unintelligible	0	0	0	0	1	0.0
Ling_Encoding	0	0	0	0	1	0.0
Style_Org	0	0	0	0	1	0.0
Style_Third	0	0	0	0	1	0.0
Style_Register	0	1	0	1	1	26.0
Style_Awkward	0	7	4	1	1	52.0
Style_Unidimoatic	2	7	1	0	1	12.0
Style_Inconsistent	0	0	0	0	1	0.0
Loc_Number	0	0	1	0	1	5.0
Loc_Currency	0	0	0	0	1	0.0
Loc_Measure	0	0	0	0	1	0.0
Loc_Time	0	0	0	0	1	0.0
Loc_Date	0	0	0	0	1	0.0
Loc_Addr	0	0	0	0	1	0.0
Loc_Tel	0	0	0	0	1	0.0
Loc_Shortc	0	0	0	0	1	0.0
AudienceAppropriateness	0	0	0	0	1	0.0
DesignMarkup	0	0	0	0	1	0.0
					Absolute Penalty Total (APT):	413.00

Evaluation Word Count (EWC):	2148	Per-Word Penalty Total (PWPT):	0.1923
Reference Word Count (RWC):	1000	Overall Normed Penalty Total (ONPT):	192.27
Penalty Scaler (PS):	1.00	Overall Quality Score (OQS):	80.77
Max. Score Value (MSV):	100.00		
		Overall Quality Fraction (OQF):	0.81
Total no. of errors	84	Sentences with errors	51.00
Total critical errors	9	Total sentences	81.00
		% Sentences with errors	0.63

Domain	MT System					
D2 ALL	ModernMT					
Error Severity Levels:	Neutral	Minor	Major	Critical	Error Type Penalty Total	
Severity Multipliers:	0	1	5	25		
Error Types	Error Counts				ET Weights	ETPTs
Term_Resource	4	6	1	0	1	11.0
Term_Inconsistent	0	0	0	0	1	0.0
Term_Wrong	0	0	0	2	1	50.0
Acc_Mistrans	1	3	10	1	1	78.0
Acc_Overtrans	0	0	0	0	1	0.0
Acc_Undertrans	0	0	0	0	1	0.0
Acc_Add	0	0	0	0	1	0.0
Acc_Omi	0	0	0	0	1	0.0
Acc_DNT	0	0	0	0	1	0.0
Acc_Untrans	0	0	0	0	1	0.0
Ling_Grammar	0	4	1	0	1	9.0
Ling_Punct	0	1	0	0	1	1.0
Ling_Spelling	0	1	1	0	1	6.0
Ling_Unintelligible	0	0	0	0	1	0.0
Ling_Encoding	0	0	0	0	1	0.0
Style_Org	0	0	0	0	1	0.0
Style_Third	0	0	0	0	1	0.0
Style_Register	0	1	0	0	1	1.0
Style_Awkward	0	9	4	1	1	54.0
Style_Unidimoatic	0	5	2	0	1	15.0
Style_Inconsistent	0	0	0	0	1	0.0
Loc_Number	0	0	0	0	1	0.0
Loc_Currency	0	0	0	0	1	0.0
Loc_Measure	0	0	0	0	1	0.0
Loc_Time	0	0	0	0	1	0.0
Loc_Date	0	0	0	0	1	0.0
Loc_Addr	0	0	0	0	1	0.0
Loc_Tel	0	0	0	0	1	0.0
Loc_Shortc	0	0	0	0	1	0.0
AudienceAppropriateness	0	0	0	0	1	0.0
DesignMarkup	0	0	0	0	1	0.0
					Absolute Penalty Total (APT):	225.00

Evaluation Word Count (EWC):	2177	Per-Word Penalty Total (PWPT):	0.1034
Reference Word Count (RWC):	1000	Overall Normed Penalty Total (ONPT):	103.35
Penalty Scaler (PS):	1.00	Overall Quality Score (OQS):	89.66
Max. Score Value (MSV):	100.00		
		Overall Quality Fraction (OQF):	0.90

Total no. of errors	58	Sentences with errors	37.00
Total critical errors	4	Total sentences	81.00
		% Sentences with errors	0.46

Domain	MT System					
D2 ALL	DeepL					
Error Severity Levels:	Neutral	Minor	Major	Critical	Error Type Penalty Total	
Severity Multipliers:	0	1	5	25		
Error Types	Error Counts				ET Weights	ETPTs
Term_Resource	6	4	1	0	1	9.0
Term_Inconsistent	0	0	0	0	1	0.0
Term_Wrong	0	0	0	0	1	0.0
Acc_Mistrans	0	1	3	1	1	41.0
Acc_Overtrans	0	0	0	0	1	0.0
Acc_Undertrans	0	3	0	0	1	3.0
Acc_Add	0	0	0	0	1	0.0
Acc_Omi	0	0	0	0	1	0.0
Acc_DNT	0	0	0	0	1	0.0
Acc_Untrans	0	0	0	0	1	0.0
Ling_Grammar	2	2	0	0	1	2.0
Ling_Punct	0	0	0	0	1	0.0
Ling_Spelling	0	0	0	0	1	0.0
Ling_Unintelligible	0	0	0	0	1	0.0
Ling_Encoding	0	0	0	0	1	0.0
Style_Org	0	0	0	0	1	0.0
Style_Third	0	0	0	0	1	0.0
Style_Register	0	0	0	0	1	0.0
Style_Awkward	0	3	0	0	1	3.0
Style_Unidimoatic	0	1	1	0	1	6.0
Style_Inconsistent	0	0	0	0	1	0.0
Loc_Number	0	0	0	0	1	0.0
Loc_Currency	0	0	0	0	1	0.0
Loc_Measure	0	0	0	0	1	0.0
Loc_Time	0	0	0	0	1	0.0
Loc_Date	0	0	0	0	1	0.0
Loc_Addr	0	0	0	0	1	0.0
Loc_Tel	0	0	0	0	1	0.0
Loc_Shortc	0	0	0	0	1	0.0
AudienceAppropriateness	0	0	0	0	1	0.0
DesignMarkup	0	0	0	0	1	0.0
					Absolute Penalty Total (APT):	64.00

Evaluation Word Count (EWC):	2215		Per-Word Penalty Total (PWPT):	0.0289
Reference Word Count (RWC):	1000		Overall Normed Penalty Total (ONPT):	28.89
Penalty Scaler (PS):	1.00		Overall Quality Score (OQS):	97.11
Max. Score Value (MSV):	100.00			
			Overall Quality Fraction (OQF):	0.97
Total no. of errors	28		Sentences with errors	19.00
Total critical errors	1		Total sentences	81.00
			% Sentences with errors	0.23

Figure 20 - MQM scorecards regarding all evaluation data, per MT engine

CROSSLANG

www.crosslang.com

CrossLang NV
Amerikagebouw Kerkstraat
106 9050 Gentbrugge
Belgium

+ 32 9 335 22 00
info@crosslang.com