Translations and Open Science

Study on machine translation evaluation
in the context of scholarly communication

D2: Outcome for discipline

"Human mobility, Environment, and Space"

Version: final

Authors:

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

Lieve Macken, Arda Tezcan (LT3)

DISCLAIMER

AVERTISSEMENT

Table of contents

1. Introduction

This deliverable outlines the evaluation outcome and best practices for the discipline “Human mobility, Environment, and Space” (ERC code SH7). 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, scores produced using automated MT metrics, examples of differences in MT output between engines, and human evaluation results.

This document is structured in the same way as D1. In Section 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 three publication types from 34 different sources of publication, and a terminology list. Table 1 gives an overview of the size and distribution.

Type of publication	Documents	Segments
Journal article	139	28812
Journal article abstract	8746	47509
Thesis abstract	3520	34120
Terminology	-	299
Total	**12405**	**110740**

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 128 additional abstracts obtained from OPERAS (hereafter referred to as the **ANR dataset**[1]). Moreover, we obtained **additional links** to bilingual abstracts and full publications from OPERAS (7 abstracts, 6 full publications – 5 of the latter also have an abstract).

Note that for the self-paced reading experiments, we strived to consider the readability level of the texts used for evaluation since they will be carried out by non-specialists in the relevant fields. However, for the SH07 discipline in particular, we did not search for additional popularising articles since the content of this discipline seems to be relatively non-technical.

2.2. Data partitioning

The dataset for SH07 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). For this discipline, we selected the ones falling under disciplines SH07 (Human Mobility, Environment, and Space) and SH03 (the Social World and Its Diversity) which are closely connected. It should be noted though that these bilingual abstracts do not constitute optimal MT training/testing material, as the translation direction is unclear and the translation is sometimes free.

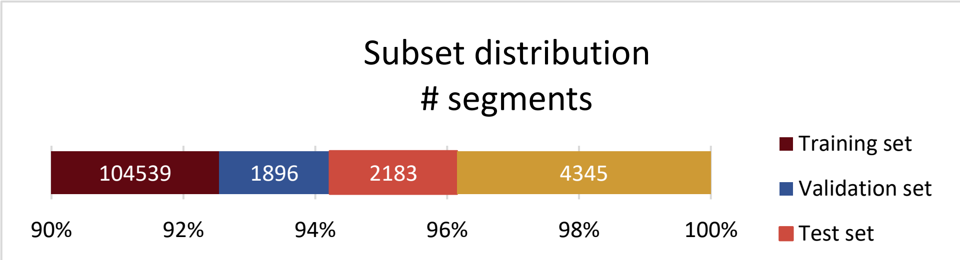


Figure 1 - Distribution of training, validation, testing, and evaluation sets

The training, validation and test sets are entirely composed of data from call 1, while the evaluation set contains both texts held out from call 1 ("internal" evaluation data) and new publications adhering to open-source copyright constraints ("external" evaluation data, see Section 2.1). The dataset partitioning aimed to ensure a balanced representation of the various text types (*journal articles*, *journal article abstracts* and *thesis abstracts*) across all subsets, as shown in Figure 2.

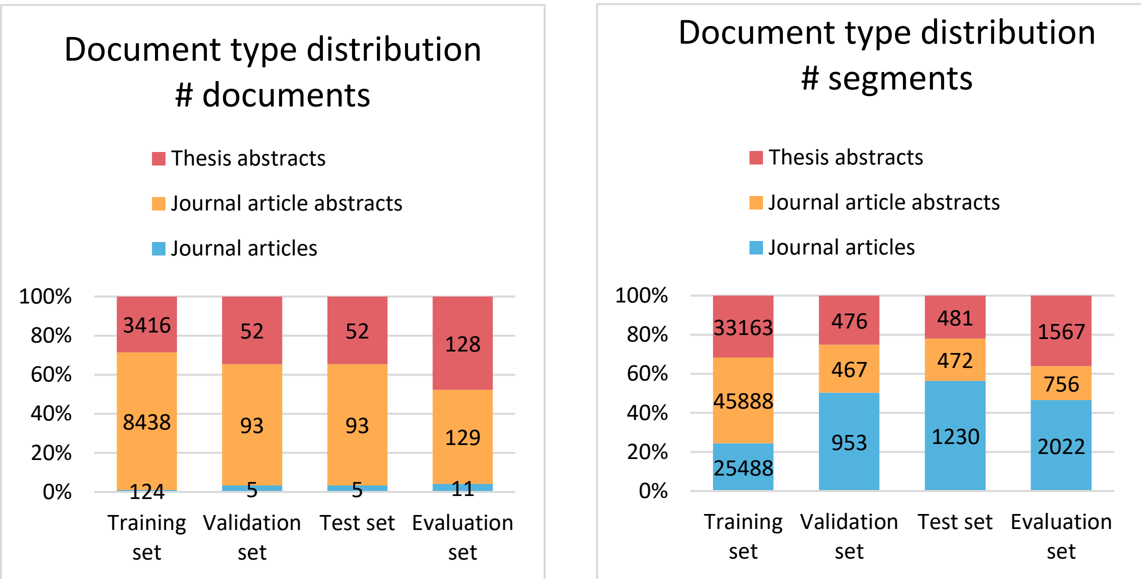


Figure 2 – Distribution of publication types for each subset, number of documents and segments

Additional information regarding the selection criteria applied to create the subsets can be found in Annex I.

2.3. MT Customisations

Table 2 gives an overview of the different operations performed. Validation set scores for the OpenNMT trainings can be found in Annex III. In addition to this, we translated the test sets using eTranslation (see Section 3) and did an OpenNMT experiment combining all data from the three disciplines (see Annex III).

Type	System	Short description[2]	Duration[3]	Date
commercial	DeepL	Baseline	/	28/03/2023
		custom (termbase)	5 seconds	28/03/2023
	ModernMT	Baseline	/	27/03/2023
		custom (OPERAS training data)	1 minute 30 seconds	27/03/2023
open source	OpenNMT	Baseline	3 h 20 m/iteration	30/03/2023
		custom 1 (OPERAS training data)	3 h 20 m/iteration	03/04/2023
		custom 2 (OPERAS training data + SciPar)	3 h 20 m/iteration	03/04/2023

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 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 performing the best. Finally, eTranslation scores are slightly lower compared to OpenNMT fine-tuned without SciPar data.

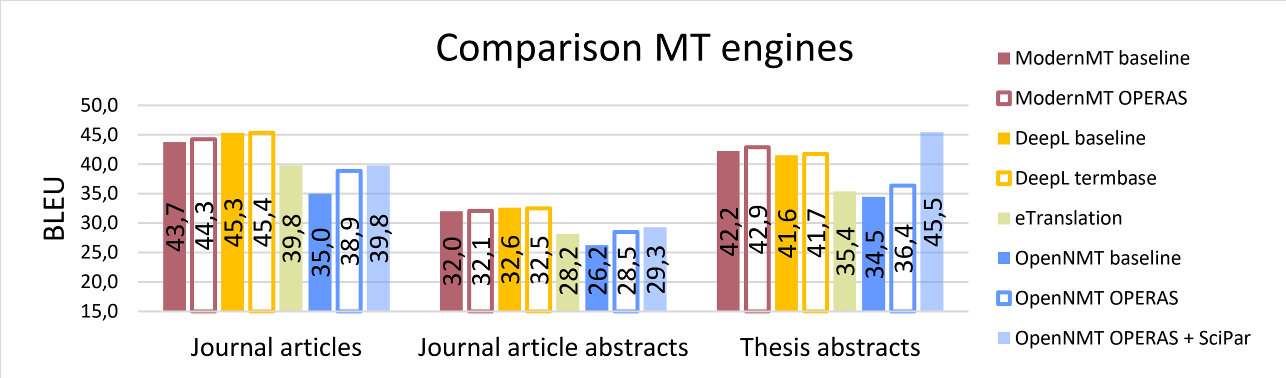


Figure 3 – Comparison of MT engines, using BLEU score, for each text type

Similar observations are made when applying other metrics (TER, ChrF, METEOR and COMET). These results are shown in Annex III:[4]

- 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.

The resulting reports with comparison view at segment level are available as separate documents and illustrated in Annex IV.

[4] In Annex III, we also provide a second set of scores for the thesis abstracts which we produced after observing a small overlap between the test set segments and SciPar, which went unobserved initially when automatically checking overlap between test set and training set because all SciPar segments end in a space. This leads the engine fine-tuned with in-domain data and SciPar to score more than 5 BLEU lower for thesis abstracts, and 1 to 2 BLEU lower for all other engines. It should be taken into consideration that DeepL and ModernMT were also potentially trained on SciPar data.

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 native speakers of English. More details about the evaluators and the feedback received can be found in Annex V.

4.2. Results of adequacy task

Based on the evaluation outcome (enriched CSV files), we produced a number of statistics and selected concrete examples showing differences in the human judgment of MT engines. For a comprehensive understanding of the adequacy task, please refer to Annex V, 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. Furthermore, the user ratings per document type indicate that journal article abstracts are less often rated as 5 (excellent). The reason may be that these types of documents contain as much information as possible, which makes them harder to translate perfectly.

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. Finally, we looked at the correlation of MT ratings between translators and researchers. Translators tend to have a higher intra-correlation than researchers.

Moreover, the intra-correlation is slightly higher compared to the inter-correlation, suggesting that translators seem to evaluate in another way than the researchers do.

4.3. Post-editing task

Based on the evaluation outcome (enriched CSV files), we produced a number of statistics and selected concrete examples showing differences in the post-edition of MT engines. These statistics are available in Annex VI. Below, we present a summary of the most interesting findings.

Post-editing times

When examining the post-editing times, we observe a large range of post-editing times, ranging 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, in terms of post-editing time, we cannot say with statistical confidence that the post-editing times differ between MT engines.

When we look at the post-editing times per document type, we see that journal article abstracts took on average the longest to edit. The difference between journal articles and thesis abstracts is smaller, although thesis abstracts took slightly shorter to edit on average.

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. Journal article abstracts on average have a perceived effort of 2.5, while journal articles and thesis abstracts only have an average perceived effort of 2.1 and 2.0 respectively.

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 corresponding to a perceived effort of 4 and 5 are different. Besides, there were just few sentences with a perceived effort of 5.

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 seems 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, as illustrated in Figure 4.

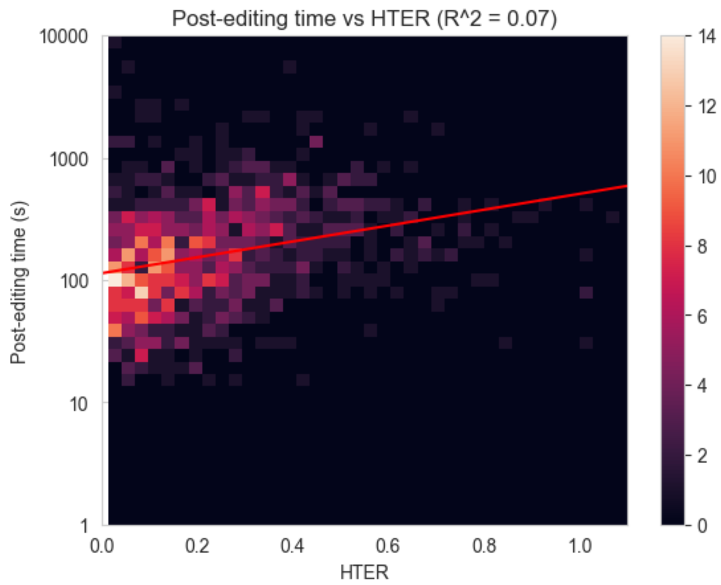


Figure 4 - Post-editing time vs HTER

4.4. Self-paced reading experiment

Data selection:

Twelve texts were selected for the discipline from three different sources: ANR thesis abstracts, full documents and TAUS thesis abstracts (see Table 3). The texts were manually selected to make sure that they were suitable for lay persons.

HUMAN MOBILITY	No. src words	No. segments
ANR - thesis abstracts		
000822_sh07_05	148	5
000750_sh03_09	152	8
000626_sh03_11	153	8
000752_sh03_11	162	9
Google docs - text excerpts full documents		
doc 5 (segments 1-9)	197	9
doc 7 (segments 1-8)	166	8
doc 8 (segments 1-7)	164	7
doc 13 (segments 1-7)	191	7
TAUS - journal abstracts		
OPERAS_000012_SH7_JAA_ZA.en	145	6
OPERAS_009241_SH7_JAA_FR.en	168	6
OPERAS_006200_SH7_JAA_FR.en	179	6
OPERAS_002064_SH7_JAA_CA.en	169	6

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

Table 4 shows the details of the full evaluation set as well as the details of the subset of segments sampled for the self-paced reading experiment. The sample was based on text type, text difficulty (manual checks), document/segment distribution, and automatic evaluation scores (BLEU, similar rankings of MT systems).[5]

	No. documents	No. segments	DeepL	OpenNMT	ModernMT
Full Evaluation Data					
thesis abstracts (ANR)	127	1566	0.43	0.37	0.40
journal articles (Google_doc)	13	931	0.55	0.40	0.48
journal article abstracts (Google_doc + TAUS)	103	1695	0.38	0.34	0.38
Subset Evaluation Data					
thesis abstracts (ANR)	4	26	0.45	0.37	0.40
journal articles (Google_doc)	4	31	0.60	0.48	0.56
journal article abstracts (Google_doc + TAUS)	4	23	0.41	0.37	0.39
TOTAL	12	80			

Table 4 - Details of the full evaluation set and subsets

The experimental design is shown in Table 5.

	SET1	SET2	SET3	SET4
ANR - thesis abstracts				
Text 1	ModernMT	OpenNMT	HT	DeepL
Text 2	OpenNMT	HT	DeepL	ModernMT
Text 3	DeepL	ModernMT	OpenNMT	HT
Text 4	HT	DeepL	ModernMT	OpenNMT
Google docs - text excerpts full documents				
Text 5	DeepL	ModernMT	OpenNMT	HT
Text 6	ModernMT	OpenNMT	HT	DeepL
Text 7	HT	DeepL	ModernMT	OpenNMT
Text 8	OpenNMT	HT	DeepL	ModernMT
TAUS - journal abstracts				
Text 9	HT	DeepL	ModernMT	OpenNMT
Text 10	ModernMT	OpenNMT	HT	DeepL
Text 11	OpenNMT	HT	DeepL	ModernMT
Text 12	DeepL	ModernMT	OpenNMT	HT

Table 5 - Experimental design for the self-paced reading experiment

The self-paced reading experiment was executed by twelve UGent staff members (aged 24-43), highly proficient in English and familiar with reading academic articles (3 participants per set).[6] Each text was presented in a cumulative way, as illustrated in Figure 5, and followed by a comprehension question and a quality assessment (see also Annex VIII).

[5] Eventually, we noticed, due to the presence of free reference translations, it may not be interesting to look at MT scores when selecting a sample to ensure these scores have a similar distribution as in the full set.

[6] All participants signed an informed consent form and got a financial reward of 10€. The experiments took place from May 2nd to May 12th 2023, with sessions lasting 30-45 minutes.

Presentation of the text (sample)

```
The increase in wild boar densities has a considerable economic cost and leads to a
 strong social tension between the stakeholders.
```

```
The increase in wild boar densities has a considerable economic cost and leads to a
 strong social tension between the stakeholders.
This tension, along with the lack of factual data, limits the possibility of
 collectively imagining other management practices.
```

Figure 5 - Text samples of the self-paced reading experiment

Comprehension question:

Is the following statement correct? The project aims to reduce human-wildlife conflict by providing factual data on wild boar.

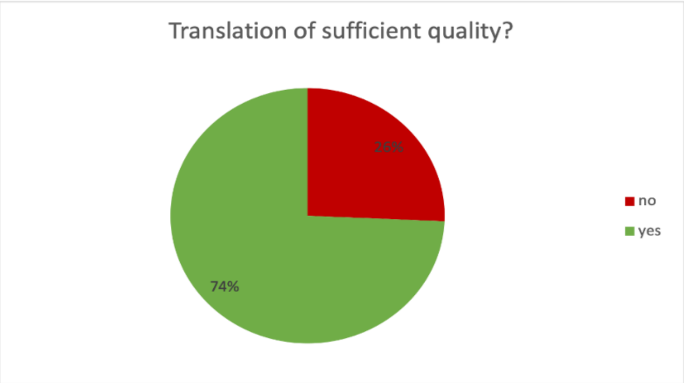
Yes/No

Quality assessment:

Was the translation of sufficient quality to get an idea of the content of the scientific text?

Yes/No

Translation quality was assessed as sufficient in 74% of all assessments. In 37 of the 144 assessments, translation quality was rated as insufficient (see Figure 6 for details).[7]



Quality Score: no		Total
	DeepL	5
	HT	8
	ModernMT	12
	OpenNMT	12
Total		**37**

Figure 6 – Sufficiency of translation quality

[7] The question why HT performs worse than DeepL requires further investigation; possible explanations may be that some human translations are free and require more reading time than literal translations and that some reference translations are produced by a (possible older type of) MT system (see Annex II).

As shown in Figure 7, average normalized reading times (milliseconds per word) were lowest for HT, i.e. human translation (463 ms) and for DeepL (467 ms), higher for ModernMT (486 ms), and highest for OpenNMT (540 ms), although there is some variation across text types (see Annex VIII).

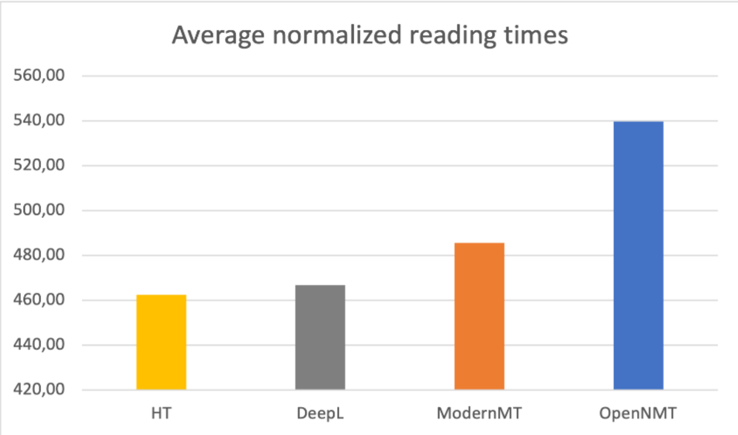


Figure 7 - Average normalized reading times

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 using the annotation platform Label Studio, see Figure 8. For a detailed description of the error annotation process, see Annex IX.

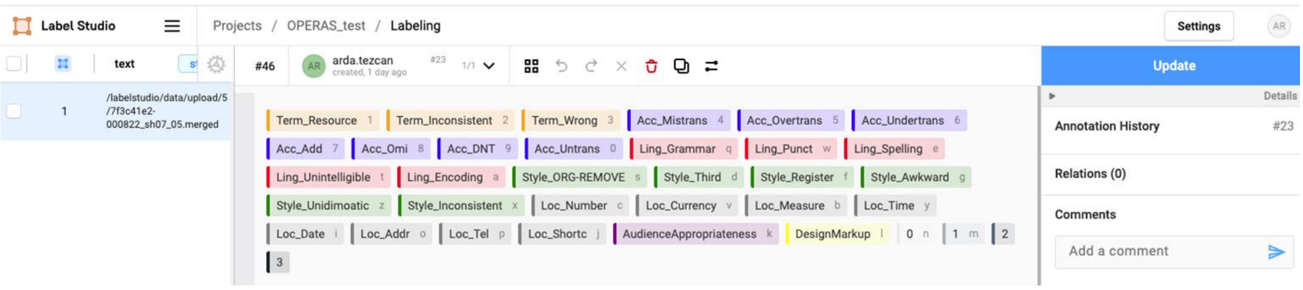


Figure 8 - Input format and taxonomy in Label Studio

After setting up and configuring LabelStudio, annotation guidelines were prepared, followed by a meeting with the evaluator and tests. Prior to error annotation, terms were marked in the source texts. The number of terms marked during both steps are as follows:

- (automatic) terms marked using the term list SH7_Mobility.tsv: 6
- (human) terms marked by the annotator: 74

Subsequently, the Label Studio files were prepared, the MT order was randomised for each file, and term annotations were transferred to Label Studio, leading to term errors to be annotated on this platform (1st priority in MQM decision tree).

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 (illustrated in Figure 9) regarding all evaluation data, per MT engine, are provided in Annex IX. In addition, we provide scorecards per text type, per engine (.xlsx) in a separate zip file (see deliverable D5). The results of other analyses are provided per text type and for the whole evaluation set, per MT engine in Annex IX and below.

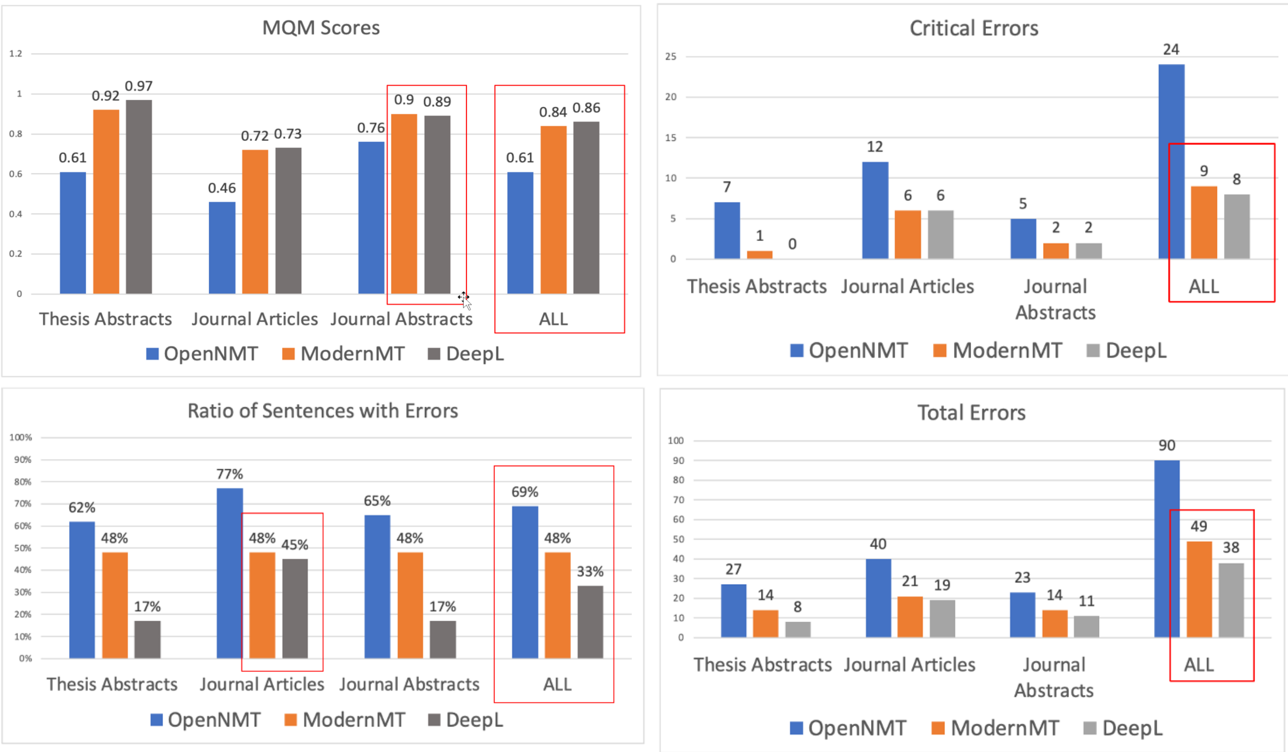


Figure 9 - MQM scorecards results

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, journal articles have a clearly higher ratio of sentences with errors than other document types in case of OpenNMT and DeepL.

5. Conclusions

In this deliverable, we presented detailed information on the first discipline "Human mobility, Environment, and Space", 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. Furthermore, the user ratings by document type indicate that journal article abstracts are less often rated as 5 (excellent). 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. Journal article 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 and between HTER and post-editing time.

From the self-paced reading experiment, translation quality was assessed as sufficient in 74% of all assessments. Average normalized reading times (milliseconds per word) were lowest for HT (463 ms) and DeepL (467 ms), higher for ModernMT (486 ms), and highest for OpenNMT (450 ms).

From the MQM scorecards and analyses, we can conclude that we obtained 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: Selection criteria for subsets

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

Training set: 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: Consists entirely of data from call 1. As we want a significant representation of each text type (journal article, journal article abstract and thesis abstract), special care needed to be taken for full journal articles, as they typically are composed of much more segments than abstracts. In order not to split up documents while still having a fair representation of different articles, a minimal number of 5 documents was used for the full articles, leading to around 1000 segments. To make sure abstracts are equally represented, we aimed to get around 500 segments for both types of abstracts. In total this leads to around 2000 segments to be separated from the training data for validation.

Test set: Same criteria as for validation set apply.

Evaluation set: To adhere to copyright constraints, three different sources were combined:

- "External" evaluation data (i.e. new publications with open-source licenses):
  - 128 open-license thesis abstracts from the **ANR dataset**
  - 7 **additional** abstracts and 6 full publications (5 of which also have an abstract)
- "Internal" evaluation data (i.e. publications held out from call 1 data):
  - All publications having an open license, resulting in a total of 198 segments.
  - Finally, in order to have a fairly representative distribution of text types in the evaluation data, we added around 500 and 1000 segments coming from (non-open-license) journal article abstracts and full journal articles respectively (no thesis abstracts from the call 1 dataset were used, as this type is already represented in the ANR dataset).

Annex II: Dataset challenges and examples

This annex gives a comprehensive 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 systematic breakdown of the various issues that arose, accompanied by relevant examples to illustrate these challenges. By doing so, we aim to shed light on the complexities, potential pitfalls and limitations when working with large datasets for machine translation.

Understanding the data

- **Machine generated reference translations**

During our dataset review, we observed that the reference translations sometimes appeared to be machine-generated. These translations were often quite literal and occasionally included errors. This was particularly noticeable in the ANR datasets. Additionally, we encountered a specific instance where an abstract explicitly mentioned that it was translated using DeepL ("Translated with www.DeepL.com/Translator (free version)").

- **Translation direction**

The translation direction was sometimes not entirely clear. Determining the correct language direction for certain datasets, particularly discerning between French and English as the source language, sometimes posed difficulties. Consequently, it is possible that, during the fine-tuning process and for the test sets, validation sets, and evaluation sets, we made use of datasets with the incorrect language direction. This could potentially influence the scores and compromise the quality of the fine-tuned data.

The original language direction is often difficult to detect, but in some cases we found the text to be human-written English translated into French (and not vice versa).

Bad source

We encountered several instances where the source text was of poor quality due to frequent errors in spelling, grammar, terminology and fluency. These mistakes adversely impacted the overall quality of the data.

Journal article abstracts:

Source FR	Reference EN
Cet article examine des inégalités dans l'accès au service de **santé medicaux maternelle** et identifie les facteurs démographiques et socio-économiques liés aux **consequences** de la **santé maternelle pauvres** en utilisant des données de cinq enquêtes démographiques et de santé conduites au Ghana (2003), au Kenya (2003), au Nigéria (2003), en Ouganda (2000-2001) et en Zambie (2001- 2002).	This paper examines inequalities in access to **maternal health care** services and identifies demographic and socio-economic factors associated with **poor maternal health outcomes** using data from five Demographic and Health Surveys conducted in Ghana (2003), Kenya (2003), Nigeria (2003), Uganda (2000-2001) and Zambia (2001-2002).

Source FR	Reference EN
Les messages dont positionné messages de planification familiale comme bénéfique pour l'individu avaient des niveaux élevés d'exposition.	Messages which positioned family planning messages as beneficial to the individual had high levels of exposure.

Source FR	Reference EN
des **Croix-tabulations** et l'analyse logistique de régression ont été employées pour évaluer l'influence des **attitudes de rôle de genre** sur le comportement sexuel risqué **tel sur-emploient** du condom et des **associés** sexuels multiples.	Cross-tabulations and logistic regression analysis were used to assess the influence of gender role attitudes on risky sexual behaviour such non-use of condom and multiple sexual partners.

Source FR	Reference EN
Étant donné la nature multidimensionnelle des facteurs prédictifs de l'exposition, **attrayante** et des messages culturellement acceptables, **grâce médiums fiables** sont susceptibles d'accroître l'exposition et attirer l'attention des hommes jeunes vers des messages de planification familiale.	Given the multivariate nature of predictors of exposure, appealing and culturally acceptable messages through reliable mediums are likely to increase exposure and attract the attention of young men towards family planning messages.

Table 6 - Bad source examples in the data

Bad reference and misalignments:

In some cases, we noticed that the reference did not fully correspond to the source text. To ensure the possibility for calculating correlations between human judgment and automatic evaluation scores, we excluded most of these cases.

Journal article abstracts:

Source FR	Reference EN
L'absence de données fiables sur l'estimation de la mortalité infantile en Afrique du Sud n'ont pas été mise à jour depuis presque dix ans à partir de 1998. Notre étude a établit les estimations sur les taux de mortalité infantile ainsi que la mortalité des enfants de moins de cinq ans.	The lack of reliable data for child mortality estimation since 1998 has meant that child mortality rates for South Africa have not been updated for almost ten years.

Table 7 - Bad reference examples in the data

Data Encoding issues

- **French accents**

The following examples taken from the TAUS datasets demonstrate instances where French accents were missing in the source text.

Journal article abstracts:

Source FR	Reference EN
Une régression binomiale négative a **identifie l'acces a léau** potable , le niveau d'éducation de la **mere** au moment de l'enquête et **à si** la **mere** est bénéficiaire d' allocation sociale pour **lénfant** comme des facteurs importants associés à la mortalité infantile.	Negative binomial regression identified the source of water, level of maternal education at the time of the survey and being a recipient of the child support grant as important factors associated with child mortality.

Source FR	Reference EN
Cependant leur effet **meme combine**, est atténué par l'immense impact du VIH qui semble avoir **submerge** les bénéfices attendus des diverses réformes de la santé.Mot clefs: **Mortalite** infantile, L'Afrique Du Sud rurale, facteurs **associes a** la **mortalite** infantile, **prevalence** du SIDA, Site de Surveillance **Demographique**.	However, their joint effect is attenuated by the overwhelming impact of HIV which also appears to have swamped the anticipated health benefit expected from various health care reforms.

Source FR	Reference EN
L'**education femine au delà** du niveau d'école secondaire ajouté aux efforts laborieux de réduire la pauvreté détiennent la **clé d'eloigner les femmes de la route vers la mort.**	Female education beyond secondary school level coupled with strenuous efforts to reduce poverty holds the key to keep women off the road to death.

Source FR	Reference EN
(p-valeur < 0,001). La prévalence du VIH dans cette **region** est parmi les plus élevées en Afrique du Sud et a augmenté de 4,2 % à 26,0 % pendant cette période, il est donc probable que l'augmentation des décès d'enfants est en grande partie attribuable à la transmission du VIH de la **mere a** l'enfant.	Maternal HIV prevalence in this area is among the highest in South Africa and rose from 4.2% to 26.0% during this period, making it probable that much of the increase in child deaths is attributable to mother to child transmission of HIV.

Table 8 - Data encoding examples in the data

Segmentation issues

- ### Sentences glued

Another issue is the frequent occurrence of multiple sentences being glued together. This problem was also present on the websites the data originated from. Consequently, the alignment between the source and reference texts did not always match accurately. This issue often resulted in incomplete translation outputs, as the machine perceived the source as a single sentence instead of multiple sentences. It also led to invalid automatic scores since the reference and source texts did not fully align with each other. We partially resolved this in the test and evaluation sets by inserting a space between sentences or splitting the sentences. However, there is still a considerable amount of glued sentences in the training data, which could also impact the quality of the fine-tuned models.

One sentence can be translated into multiple sentences and vice versa.

Journal article abstracts:

Source FR	Reference EN
L'analyse des **table de vie** révèle une inversion de la tendance à la baisse du taux de mortalité entre 1990 et 2000 **.,** Pendant cette **periode**, la mortalité infantile a augmenté de 43 à 65 par 1000 naissances et celle des moins de 5 ans de 65 à 116 pour 1 000 naissances, ce qui se traduit par un **RR** de 1,85 pour la **periode etudiee**.	Life table analysis of the data reveals a reversal of the downward trend in mortality rates over time that began around 1990 in this population**.**Between 1990 and 2000 infant mortality increased from 43 to 65 per 1000 live births and under-five mortality from 65 to 116 per 1 000 live births which translates into a RR of 1.85 over the 10 year period (p-value <0.001).

Table 9 - Glued sentences in the data

- **Words glued**

Not only were sentences glued together, but there were also instances where a space was missing between certain words. Again, an additional space was usually added in the test and evaluation sets.

Journal article abstracts:

Source FR	Reference EN
Cellesci se basent principalement sur les connaissances des experts et leurs arguments qui permettent d'illustrer les incertitudes liées aux évolutions démographiques futures.	It **dependsmainly** on experts' knowledge and arguments which can help to illustrate the uncertainties associated with future demographic trends.

Table 10 - Glued words in the data

Freely translated outputs

Additionally, we observed that certain sentences were quite freely translated. This could potentially impact the automatic scores considering source and reference for evaluating MT outputs.

Evaluation Setup Challenges

- **OpenNMT/ModernMT/DeepL: part of translation missing**

As mentioned earlier, the issue of glued sentences in the source text could result in missing parts of translations. To address this concern, whenever we identified such cases, we inserted an additional space in the source text and re-generated all machine translation outputs.

Annex III: Automatic scores

Table 11 and Table provide metric scores for all document types. Table 3 provides validation scores. Table 4 shows the automatic scores for the OpenNMT training with all data from the three disciplines.

Automatic scores

		30 epochs (OpenNMT)					60 epochs (OpenNMT)		
Type	**Engine**	**SacreBLEU**	**TER**	**METEOR**	**ChrF**	**COMET**	**SacreBLEU**	**TER**	**METEOR**
Journal articles	ModernMT baseline	43,74	43,7	35,6	67,92	85,77	/	/	/
	ModernMT OPERAS	44,27	43,8	35,7	68,09	85,71	/	/	/
	Deepl baseline	45,31	42,8	36,2	68,88	85,61	/	/	/
	Deepl termbase	45,35	42,8	36,2	68,9	86,61	/	/	/
	eTranslation	39,77	47,12	33,4	64,73	83,77	/	/	/
	OpenNMT baseline	35,02	51,2	30,8	61,53	81,75	36,12	50,8	31,6
	OpenNMT OPERAS	38,87	48,1	33,1	64,67	82,9	38,49	48,8	32,9
	OpenNMT OPERAS + SciPar	39,81	47,5	33,6	65,35	83,82	39,54	47,8	33,3
Journal article abstracts	ModernMT baseline	32	56,3	29,4	61,14	83,81	/	/	/
	ModernMT OPERAS	32,08	56,7	29,7	61,24	83,7	/	/	/
	Deepl baseline	32,61	55,6	29,8	61,74	84,52	/	/	/
	Deepl termbase	32,49	55,6	29,8	61,72	84,49	/	/	/
	eTranslation	28,18	59,34	27,6	58,24	82,55	/	/	/
	OpenNMT baseline	26,23	61	26,3	56,51	81,24	26,59	61,2	26,6
	OpenNMT OPERAS	28,51	58,8	27,5	58,32	81,73	27,84	59,1	27,2
	OpenNMT OPERAS + SciPar	29,28	58,1	27,8	58,99	82,57	29,03	58,3	27,8

Table 11 - Automatic scores for journal articles and their abstracts

		30 epochs (OpenNMT)					60 epochs (OpenNMT)		
Thesis abstracts	ModernMT baseline	42,22	46,3	35	68,33	85,39	/	/	/
	ModernMT OPERAS	42,9	46,3	35,3	68,65	85,3	/	/	/
	Deepl baseline	41,55	47,2	34,9	68,04	85,67	/	/	/
	Deepl termbase	41,73	47	34,9	68,09	85,66	/	/	/
	eTranslation	35,36	51,82	32	63,38	83,4	/	/	/
	OpenNMT baseline	34,46	52,6	31,5	62,73	82,2	34,74	52,6	31,8
	OpenNMT OPERAS	36,38	51,4	32	64,32	83,11	36,37	51,1	32,1
	OpenNMT OPERAS + SciPar	45,46	44,2	35,8	69,22	84,87	47,74	42,6	36,8
Thesis abstracts, filtered	*ModernMT baseline*	*39,91*	*49,2*	*33,2*	/	/	/	/	/
	ModernMT OPERAS	*41,03*	*49*	*33,5*	/	/	/	/	/
	Deepl baseline	*40,47*	*49,3*	*33,6*	/	/	/	/	/
	Deepl termbase	*40,6*	*49,2*	*33,6*	/	/	/	/	/
	OpenNMT baseline	/	/	/	/	/	*33,49*	*54,1*	*30,5*
	OpenNMT OPERAS	/	/	/	/	/	*34,98*	*53,1*	*31*
	OpenNMT OPERAS + SciPar	/	/	/	/	/	*42,13*	*47,7*	*34*

Table 12 - Automatic scores for thesis abstracts

Validation scores

Validation set	**OpenNMT**	**SacreBLEU**				
Engine	**10 epochs**	**20 epochs**	**30 epochs**	**40 epochs**	**50 epochs**	**60 epochs**
OpenNMT baseline	32,20	32,30	33,10	33,40	34,20	34,00
OpenNMT OPERAS	34,20	34,30	34,60	34,50	35,10	35,10
OpenNMT OPERAS + SciPar	36,20	37,40	37,40	37,60	37,30	37,50

Table 13 - BLEU score on validation set for every 10 iterations

We noticed that, in case of thesis abstracts, using SciPar as training data for OpenNMT leads to a much higher score than the baseline (+ 13 BLEU after 60 epochs in block "thesis abstracts non-filtered" in Table). We checked the SciPar data and it appears there is some overlap with the thesis abstracts used as test data (159 of the 2183 test segments occur in SciPar), due to a small issue (all segments in SciPar end in a space, so our automatic comparison of train vs. test data did not detect the overlap). After filtering out the overlapping segments from the test data, we recalculated the scores (see block "thesis abstracts filtered"). This leads to lower scores for all engines than in the preceding block.

Training with data from all disciplines combined

		20 epochs (ALL)		
Type	**Engine**	**SacreBLEU**	**TER**	**METEOR**
TA	OpenNMT baseline	34,46	52,6	31,5
	OpenNMT OPERAS + SciPar (SH7)	45,46	44,2	35,8
	OpenNMT OPERAS + SciPar (ALL)	40,63	47,5	34
JAA	OpenNMT baseline	26,23	61	26,3
	OpenNMT OPERAS + SciPar (SH7)	29,28	58,1	27,8
	OpenNMT OPERAS + SciPar (ALL)	30,83	57	28,7
JA	OpenNMT baseline	35,02	51,2	30,8
	OpenNMT OPERAS + SciPar (SH7)	39,81	47,5	33,6
	OpenNMT OPERAS + SciPar (ALL)	41,97	45,7	34,6

Table 14 - Automatic scores when training with data of all disciplines combined

Annex IV: Automatic report examples

By means of illustration, we show examples of both improving and decreasing quality after fine-tuning the different models. These examples derive from the test set for automatic evaluation.

DeepL baseline versus customised

Figure 10 provides a rare example of a term being corrected according to the glossary uploaded. Most of the time, terms were already correctly translated by the baseline. Figure 11 shows such a term (“professional mobility”). In this example, the custom model also erroneously transforms another term, “specialised educators” into “social workers”, which is not a termbase entry.

Type	Sentence	dist	BLEU
SRC	La réceptivité habitante à l'épreuve des projets d'habitat social : enjeux et perspectives à travers le cas de Marseille : la rénovation urbaine à Saint-Barthélemy III Picon-Busserine.	-	-
REF	The resident receptivity proof of social housing projects: Challenges and prospects through the case of Marseille: urban renewal in St. Bartholomew III Picon-Busserine.	-	-
deepl baseline	Resident receptivity to the test of social housing projects: issues and perspectives through the case of Marseilles: the urban renovation of Saint-Barthélemy III Picon-Busserine.	51	0.2801
REF/deepl baseline	~~R~~The resident receptivity ~~to the test~~proof of social housing projects: ~~issu~~Challenges and p~~e~~rospect~~ive~~s through the case of Marseille~~s: the~~: urban ren~~ovation of Saint-Barthélemy~~ewal in St. Bartholomew III Picon-Busserine.	51	0.2801
deepl	Resident receptivity to the test of social housing projects: issues and perspectives through the case of Marseille: the urban renewal in Saint-Barthélemy III Picon-Busserine.	43	0.40299999999999997
REF/deepl	~~R~~The resident receptivity ~~to the test~~proof of social housing projects: ~~issu~~Challenges and p~~e~~rospect~~ive~~s through the case of Marseille~~: the~~ urban renewal in S~~aint-Barthélemy~~t. Bartholomew III Picon-Busserine.	43	0.40299999999999997

Figure 10 - Rare example of term (“urban renewal”) corrected according to the glossary uploaded

Type	Sentence	dist	BLEU
deepl baseline	Transferability of knowledge and skills in the training and professional mobility of specialised educators in the European area. Comparative study between Italy and France	25	0.8556999999999999
REF/deepl baseline	Transferability of knowledge and skills in the training and professional mobility of ~~specialised educators~~educators specializing in the European area. Comparative study between Italy and France	25	0.8556999999999999
deepl	Transferability of knowledge and skills in the training and professional mobility of social workers in the European area. Comparative study between Italy and France	23	0.8467
REF/deepl	Transferability of knowledge and skills in the training and professional mobility of ~~social workers~~educators specializing in the European area. Comparative study between Italy and France	23	0.8467
SRC	Transférabilité des savoirs et des compétences dans la formation et la mobilité professionnelle des éducateurs spécialisés dans l'espace européen. Étude comparative entre l'Italie et la France	-	-
REF	Transferability of knowledge and skills in the training and professional mobility of educators specializing in the European area. Comparative study between Italy and France	-	-

Figure 11 - Examples of terms translated well by baseline but not necessarily by custom model

ModernMT baseline versus customised

Figure 12 shows an example of corrections made by the custom ModernMT model: locale (*spatialized* > *spatialised*) and subclause *what some people find just, others find completely injust*. The latter has a match in the training data: *In Tunisia, for example, Fautras writes of "the subjective and spatialised dimension of injustice: what some people find just, others find completely unjust."*. Note that the locale used (US versus UK English) is not consistent throughout the outputs (baseline and custom). Figure 13 shows another example of changes done by the custom ModernMT model: "ramassages manuels ou mécaniques" receives a more literal translation "mechanical pick-ups".

Type	Sentence	dist	BLEU
SRC	En questionnant la façon dont la contestation organisée par Salah est perçue par les habitants de la région de Regueb, on peut mesurer la dimension subjective et spatialisée de l'injustice : ce qui paraît juste aux uns peut sembler tout à fait injuste à d'autres.	-	-
REF	By looking at how the inhabitants of the Regueb region perceive the opposition organised by Salah, we can measure the subjective and spatialised dimension of injustice: what some people find just, others find completely unjust.	-	-
modernmt baseline	By questioning the way in which the protest organized by Salah is perceived by the inhabitants of the Regueb region, one can measure the subjective and spatialized dimension of injustice: what seems fair to some may seem quite unfair to others.	144	0.2915
REF/modernmt baseline	By ~~questioning the way in which the protest organized by Salah is~~looking at how the inhabitants of the Regueb region perceive~~d by~~ the ~~inhabitants of the Regueb region, or~~opposition organised by Salah, we can measure the subjective and spatiali~~z~~sed dimension of injustice: what s~~eems fair to some may seem quite unfair to others~~ome people find just, others find completely unjust.	144	0.2915
modernmt	By questioning the way in which the protest organized by Salah is perceived by the inhabitants of the Regueb region, one can measure the subjective and spatialised dimension of injustice: what some people find just, others find completely unjust.	94	0.6069
REF/modernmt	By ~~questioning the way in which the protest organized by Salah is~~looking at how the inhabitants of the Regueb region perceive~~d by~~ the ~~inhabitants of the Regueb region, or~~opposition organised by Salah, we can measure the subjective and spatialised dimension of injustice: what some people find just, others find completely unjust.	94	0.6069

Figure 12 - Corrections made by the custom ModernMT model

Type	Sentence	dist	BLEU
SRC	Quels sont les impacts des ramassages manuels ou mécaniques de ces laisses sur le bilan sédimentaire et la dynamique des plages?	-	-
REF	What are the impacts of the manual or mechanical collection of these drift materials on the sediment budget and dynamics of beaches?	-	-
modernmt baseline	What are the impacts of manual or mechanical collection of these leashes on the sediment balance and beach dynamics?	40	0.4335
REF/modernmt baseline	What are the impacts of the manual or mechanical collection of these ~~leashe~~drift materials on the sediment b~~alance and beach dynamic~~udget and dynamics of beaches?	40	0.4335
modernmt	What are the impacts of manual or mechanical pick-ups of these leashes on the sediment balance and beach dynamics?	51	0.2871
REF/modernmt	What are the impacts of the manual or mechanical ~~pick-ups of these leashes on the sediment balance and beach dynamic~~collection of these drift materials on the sediment budget and dynamics of beaches?	51	0.2871

Figure 13 - More literal translation by custom MT model

OpenNMT baseline versus customized

Type	Sentence	dist	BLEU
SRC	C'est une étude sur l'appropriation et la construction de ce système en fonction des actions menées par les usagers, les ingénieurs, les législateurs, les clubs automobiles, les services de voirie ou les organes de l'administration routière.	-	-
REF	This is a study focusing on the appropriation and construction of this system through the interventions of users, engineers, legislators, automobile clubs, road services and administration.	-	-
OpenNMT baseline	It is a study of ownership and construction of this system in the light of actions by users, engineers, legislators, car clubs, street services or road administration bodies.	73	0.322
REF/OpenNMT baseline	~~It~~This is a study ~~of ownership~~focusing on the appropriation and construction of this system ~~in the light of ac~~through the interventions ~~by~~of users, engineers, legislators, ~~car~~automobile clubs, ~~street~~road services ~~or ro~~and administration ~~bodies~~.	73	0.322
OpenNMT	It is a study on the appropriation and construction of this system based on the actions undertaken by users, engineers, legislators, car clubs, road services or road administration bodies.	63	0.4551
REF/OpenNMT	~~It~~This is a study focusing on the appropriation and construction of this system ~~based on the actions undertaken by~~through the interventions of users, engineers, legislators, ~~car~~automobile clubs, road services ~~or ro~~and administration ~~bodies~~.	63	0.4551
OpenNMT SciPar	This is a study on the appropriation and construction of this system based on the actions led by users, engineers, legislators, automobile clubs, road services or the organs of road administration.	50	0.5558
REF/OpenNMT SciPar	This is a study focusing on the appropriation and construction of this system ~~based on the actions led by~~through the interventions of users, engineers, legislators, automobile clubs, road services ~~or the organs of ro~~and administration.	50	0.5558

Figure 14 - OpenNMT changes by customised models

Figure 14 shows an example of OpenNMT improvements: terms such as “automobile clubs” and “road services” get correctly translated. Note that at the same time “road administration bodies” (baseline) is transformed into “organs of road administration” (custom models), which can be considered as a poor translation.

Annex V: Adequacy task

Setup and execution

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, 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 native speakers of English. We decided to reduce the envisaged number of segments from the planned 500 per task to 400 for time and budget reasons, and proposed a price to the evaluators and a time span of two weeks for performing the work. 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). 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.

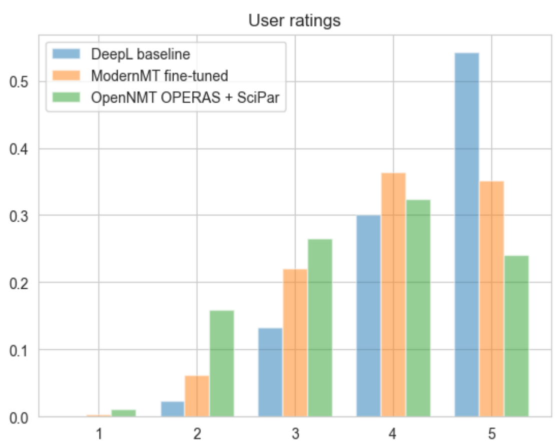
Some of the evaluators provided feedback relating to the tasks:

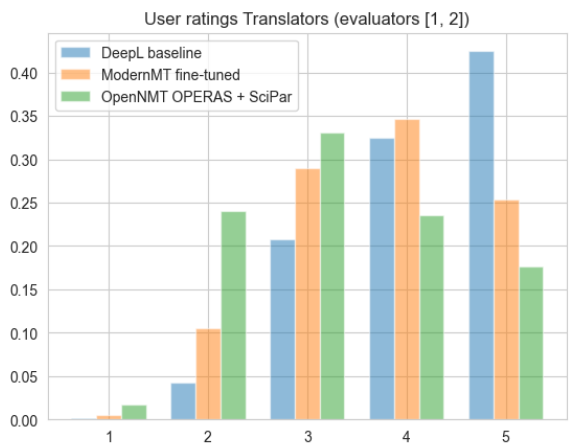
- One translator commented on the evaluation scheme in the instructions: "In the adequacy task, the difficulty was getting used to the evaluation grid itself, given that a segment could be categorised as "excellent" in terms of MT adequacy (i.e. understandable without reference to the ST and using the appropriate terminology) even though it might be barely adequate in terms of fluency and readability (which were not included in the criteria). The grid also makes the distinction between "all", "most", "much" and "little" meaning, which does not take into account the criticality of the meaning errors, so that for instance, "most" of the meaning may be conveyed, but the key point would be missed or misunderstood by the TT reader. Other difficulties were often related to the quality of the ST, i.e. the TT was not immediately understandable because the ST was unclear or ambiguous. So should the TT be classed as "excellent" because it perfectly conveys the inadequacies of the ST?"
- Another translator noticed that, when going back in MT-Eval to an already rated segment, the order of the MT outputs changed with respect to the previously shown order (when the user moves to a segment, the tool automatically orders the outputs randomly). Some of the comments the translator provided in the dedicated field in MT-Eval included the number of the MT output being commented on. We explained to the evaluator that a comment should be clear on the MT output being commented on. In practice, this already appeared to be the case in the comments of the evaluators, so the numbers can be ignored.

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 15 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.





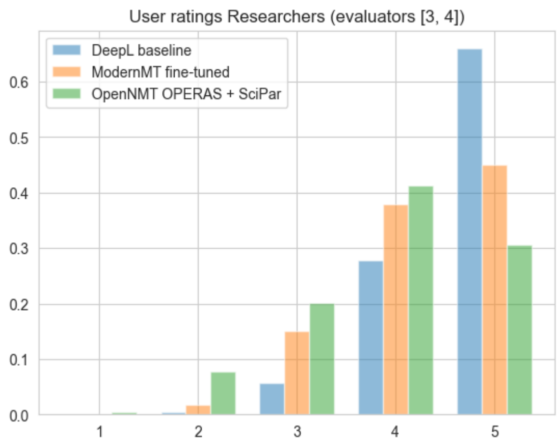


Figure 15 - User ratings per type of evaluator

Figure 16 shows the distribution of all evaluators' ratings per document type. We cannot say with significant confidence that the average rating differs between the types. While the mean value is similar, the distribution slightly differs. It seems that journal article abstracts are less often rated as 5 (excellent). It seems that it is harder for these documents to get the highest rating, possibly due to the fact that these types of documents contain as much information as possible, which makes them harder to translate perfectly. When checking the distributions separately per type of evaluator, we came to the same conclusions as in case no distinction was made.

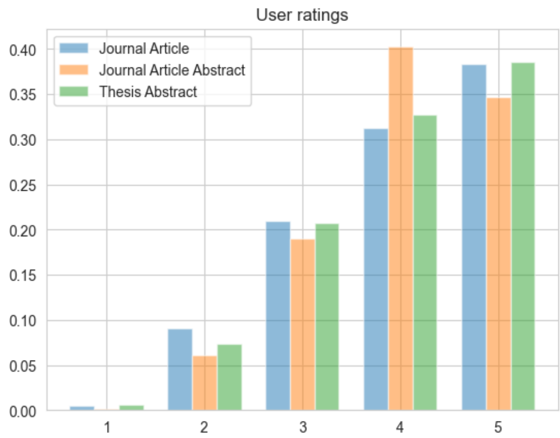


Figure 16 - User ratings per document type

Another statistic we produced relates to the MT engine rankings implicitly assigned by evaluators through the ratings they provided. This is shown in Figure 17, 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 18, 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. Looking at the correlation of MT ratings between translators and researchers, shown in Figure 19, we observe that translators tend to have a higher intra-correlation than researchers. Moreover, the intra-correlation is slightly

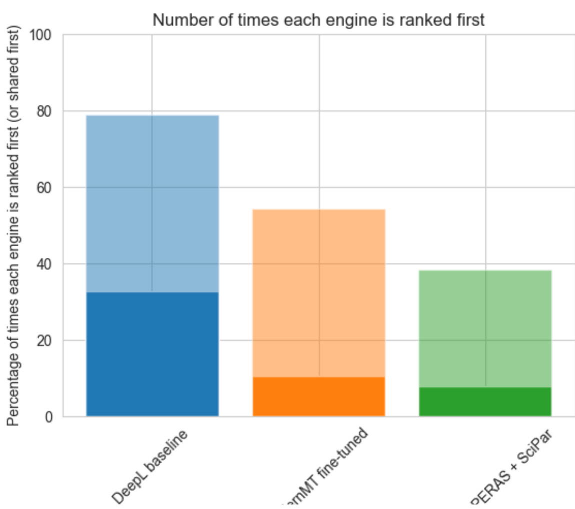
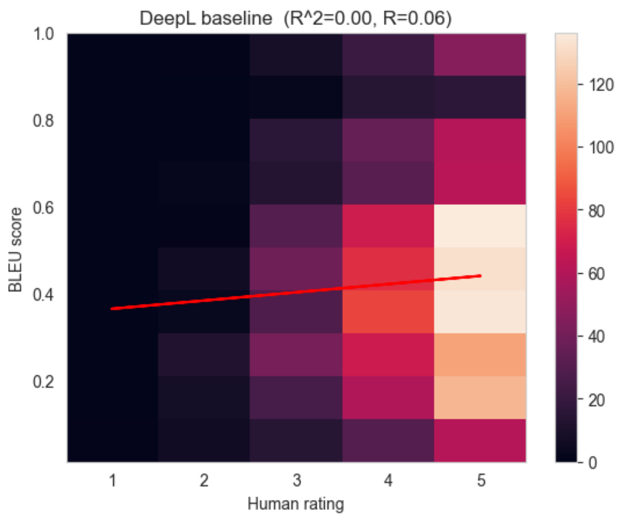
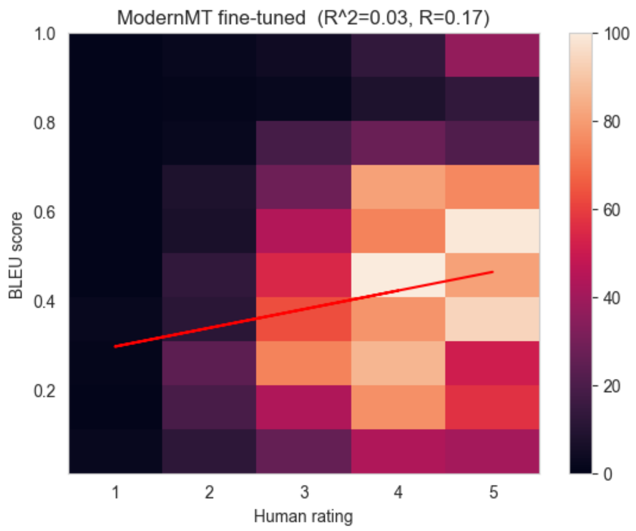
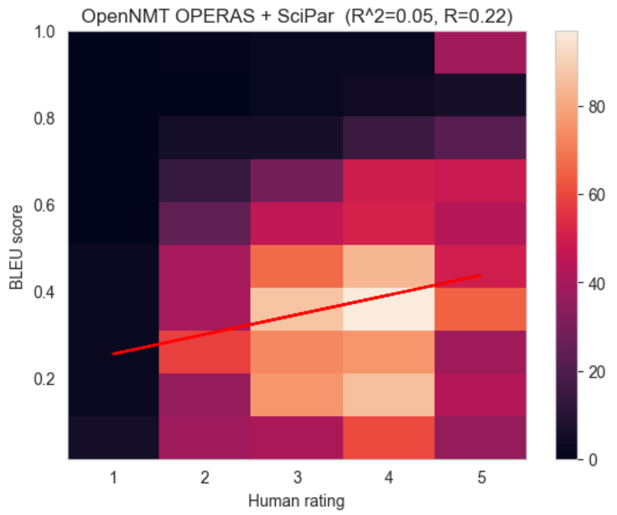


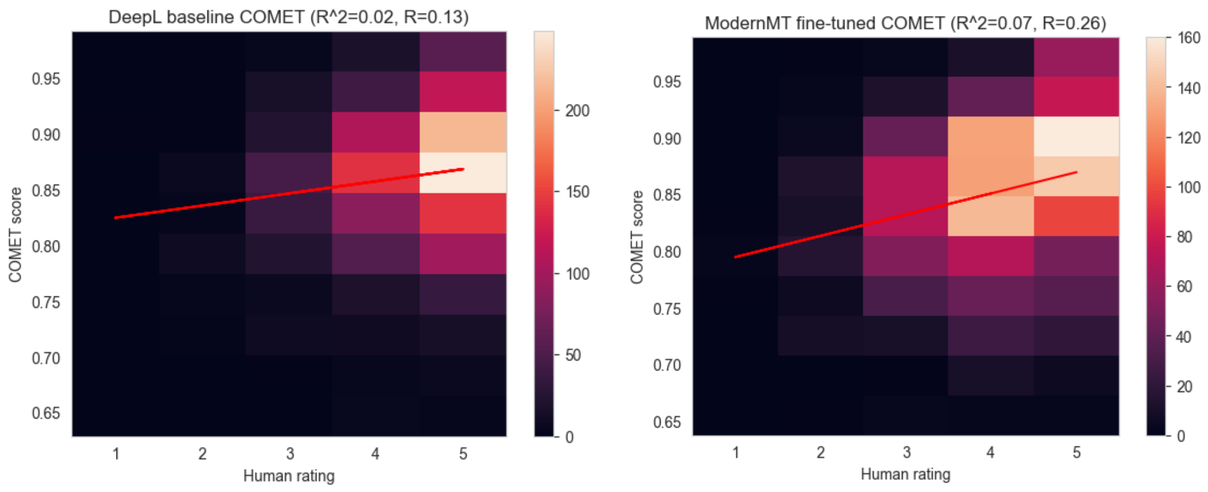
Figure 17 - Number of times engines are ranked first

higher compared to the inter-correlation, suggesting that translators seem to evaluate in another way than the researchers do.









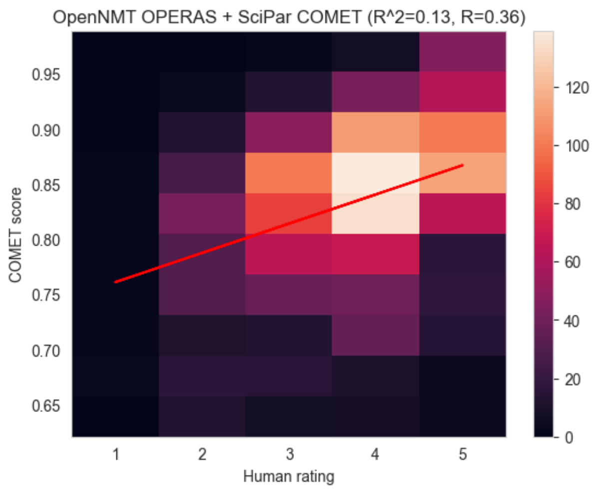


Figure 18 - Correlation between automatic metrics and human ratings

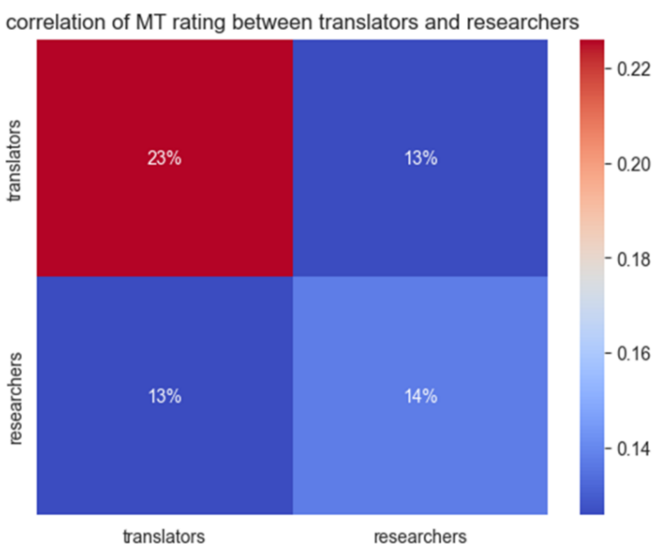


Figure 19 - Correlation of MT rating between translators and researchers

Annex VI: 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 as those executing the adequacy task, and by two researchers native speakers of French. We decided to reduce the envisaged number of segments from the planned 500 per task to 400 for time and budget reasons, and proposed a price to the evaluators and a time span of two weeks for performing the work. Evaluators were paid by the hour. 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 20 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, ranging 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.

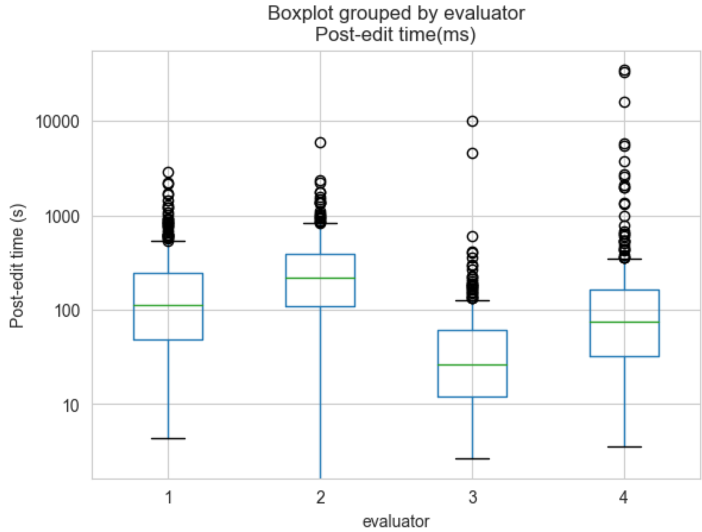


Figure 20 - Boxplot grouped by evaluator - post-edit time (ms)

When investigating the correlation between post-edit time and perceived effort, we obtain Figure 21. It shows the median post-edit time together with a confidence interval of the median. 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. There were also just few sentences with a perceived effort of 5.

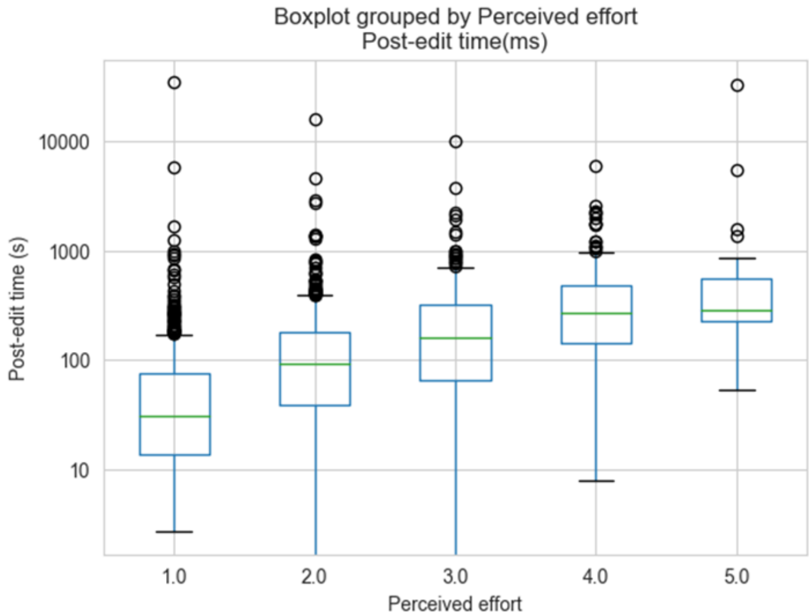


Figure 21 - Boxplot grouped by perceived effort - post-edit time (ms)

Figure 22 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 in terms of post-edit time, we cannot say with statistical confidence that the post-edit times differ between MT engines.

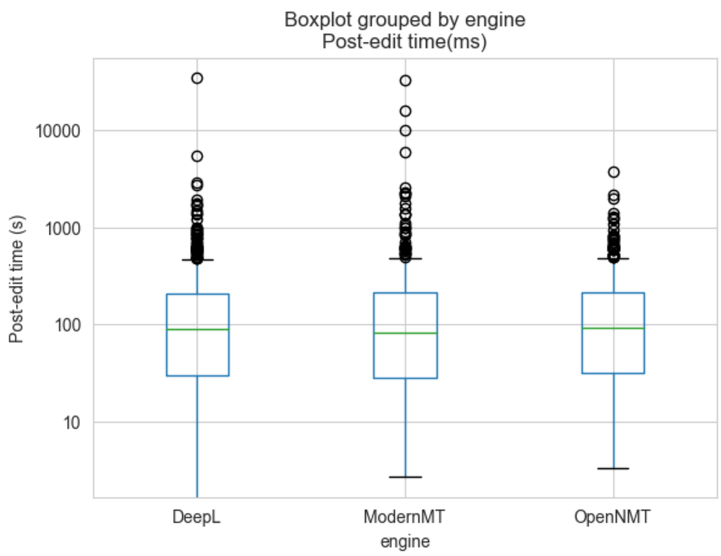


Figure 22 - Boxplot grouped by engine - post-edit time(ms)

If we look at the different evaluators we get the results in Table 6. The ranking differs among the evaluators. This confirms that we cannot clearly distinguish the engines in terms of post-editing time.

	DeepL (median post-edit time +- SEM)	**ModernMT**	**OpenNMT**
Translator 1	[114 s, 141 s]	[91 s, 112 s]	[102 s , 126 s]
Translator 2	[191, 228]	[239, 295]	[190, 224]
Researcher 1	[25, 30]	[21, 25]	[26, 31]
Researcher 2	[75, 93]	[58, 74]	[65, 80]

Table 6 - Post-edit time grouped by evaluator and engine

In Figure 23, 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.

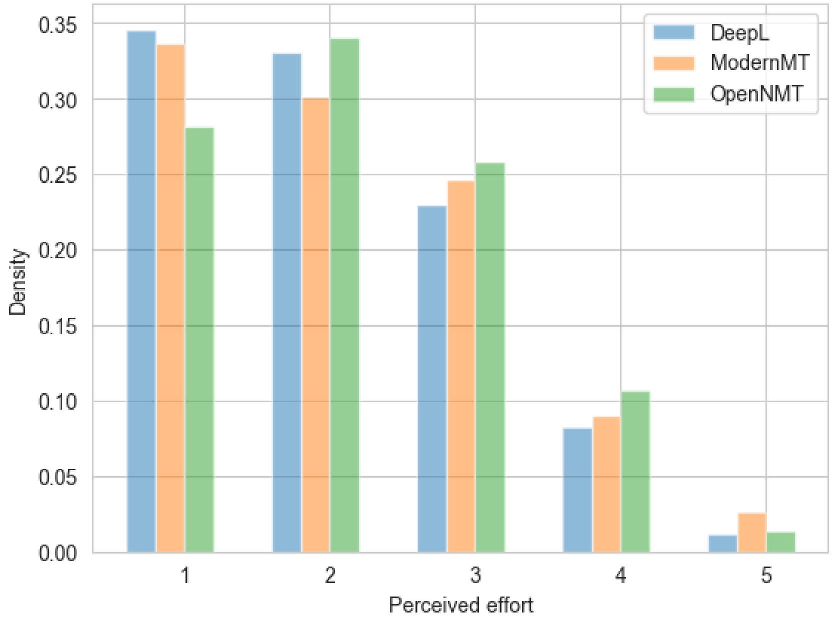


Figure 23 – Perceived effort per engine

Figure 24 shows the post-editing time per document type. The journal article abstracts took on average the longest to edit. The difference between journal articles and thesis abstracts is smaller, although thesis abstracts took slightly shorter to edit on average.

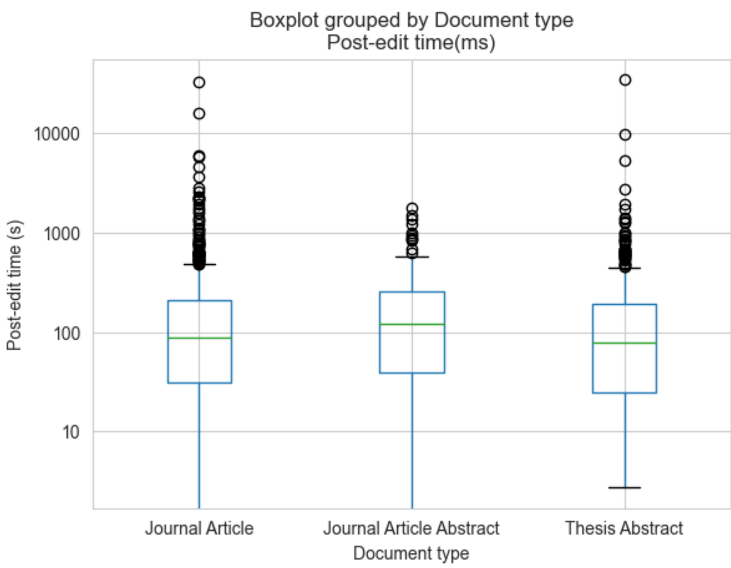


Figure 24 - Boxplot grouped by document type - post-edit time (ms)

The comparison of perceived efforts in Figure 26 confirms the previous findings. The journal article abstracts on average have a perceived effort of 2.5, while the journal articles and thesis abstracts only have a average perceived effort of 2.1 and 2.0 respectively.

When calculating the HTER and comparing it with the perceived effort, we can clearly see a correlation, as shown in Figure 25. While the median HTER of perceived effort 5 seems to be lower than for perceived effort 4 (Figure 27), we have too few samples to make any significant conclusions for this.

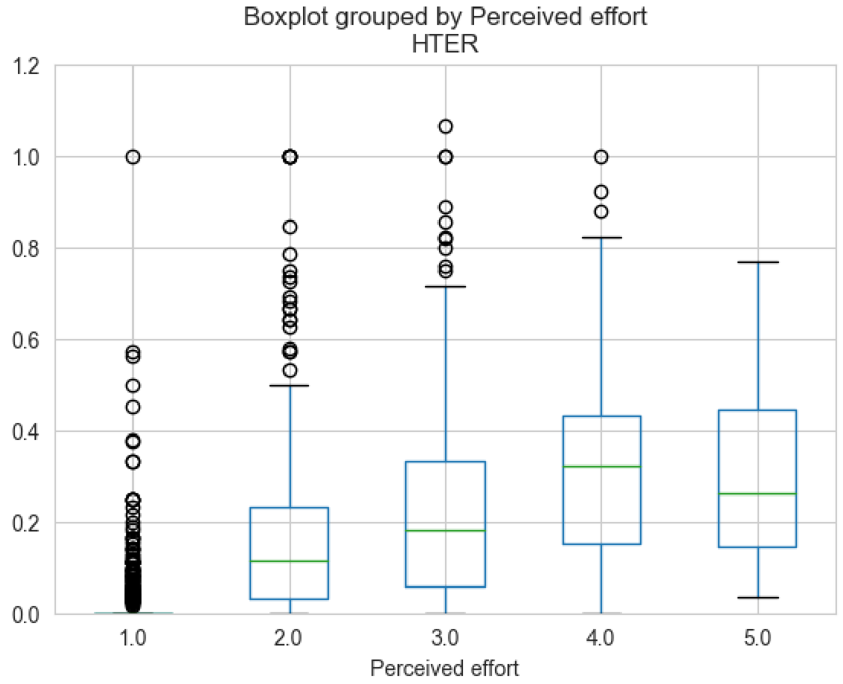


Figure 25 - Boxplot grouped by perceived effort - HTER

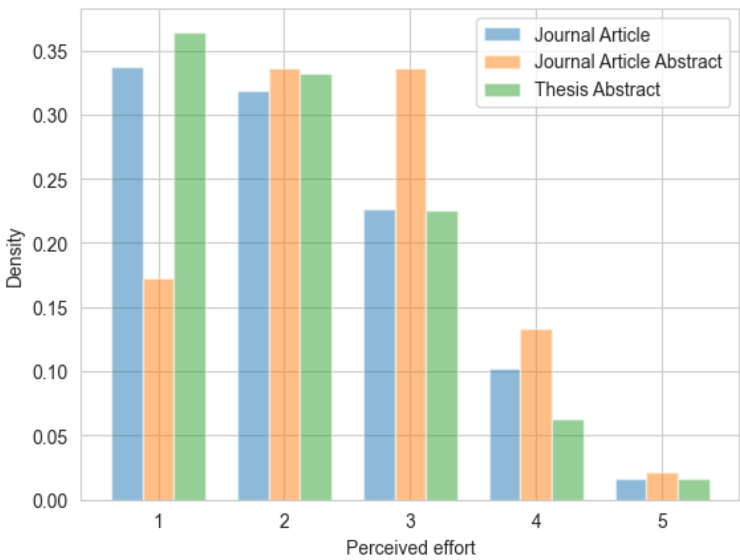


Figure 26 - Perceived effort per document type

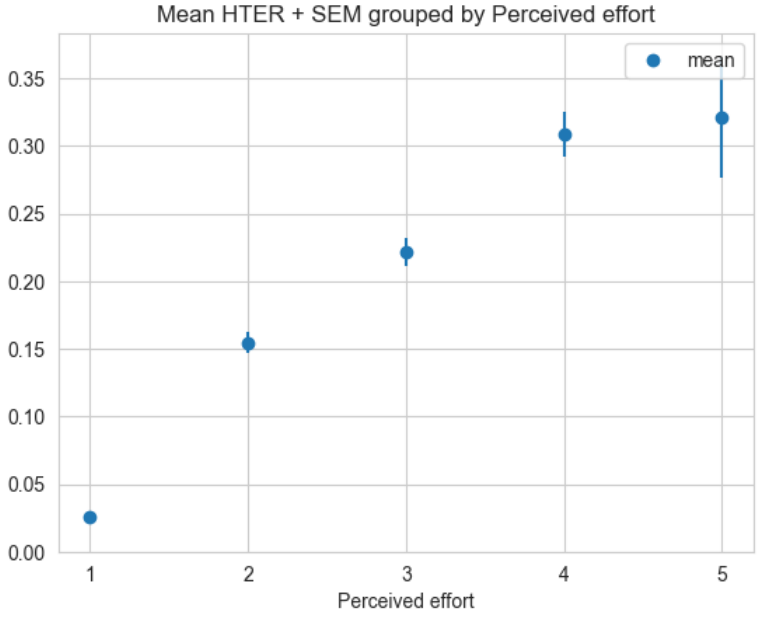


Figure 27 - Mean HTER + SEM grouped by perceived effort

There is a correlation between post-editing time and HTER, as illustrated in Figure 28.

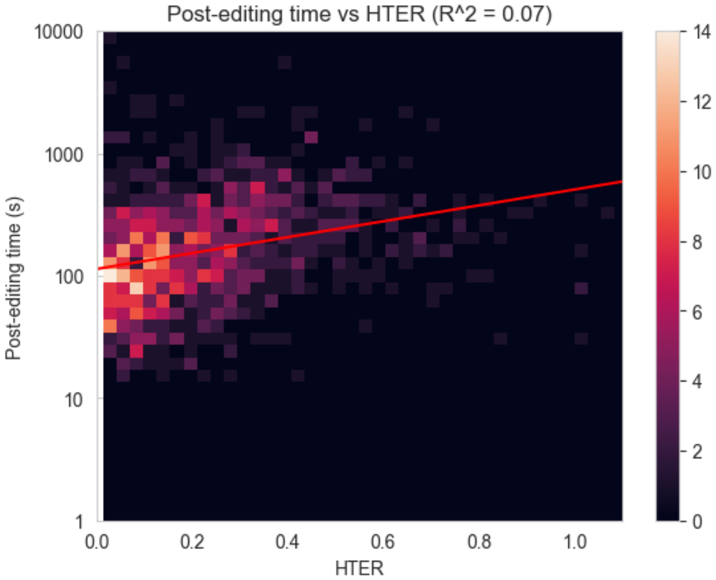


Figure 28 - Post-editing time vs HTER

Annex VII: Human evaluation examples

These examples derive from the human evaluation set.

Adequacy

Differences between engines

The segment numbers relate to the position in the MT-Eval input.

Segment	Evaluator	Source	DeepL baseline	ModernMT fine-tuned	OpenNMT Operas + SciPar
18		Il semblerait aussi que dans les régions plus éloignées où l'implantation des nouveaux arrivants est moins importante, et où la question du rapport interculturel avec les settlers , les colons, comme on les appelle en anglais, ou la settler society , la société colonisatrice, se pose moins, les négociations se règlent plus facilement.	It would also seem that in more remote regions where the settlement of newcomers is less important, and where the question of the intercultural relationship with the settlers, or the settler society, is less of an issue, negotiations are more easily resolved.	It would also seem that in more remote regions, where the settlement of newcomers is less important, and where the question of the intercultural relationship with settlers , settlers, as they are called in English, or the settler society , the colonizing society, arises less, negotiations are more easily settled.	It would also seem that in the more remote regions where the settlement of newcomers is less important, and where the question of the intercultural relationship with the settlers, the settlers, as is called in English, or settler society, the colonizing society, is less likely to settle down.
	Evaluator 1		5 -Excellent	2 - Poor	2 - Poor
	Evaluator 2		5- Excellent	2 - Poor	2 - Poor
	Evaluator 3		5 - Excellent	2 - Poor	2 - Poor
	Evaluator 4		4 - Good	3 - Fair	3 - Fair
23		Ces taux seront ensuite intégrées aux modèles structurels pour en déduire la proportion de postes fantômes, la distribution des efforts de recherche par âge des annonces, la distribution des âges de renouvellement des annonces et l'impact des informations obsolètes sur le chômage et l'efficacité du processus d'appariement.	These rates will then be incorporated into structural models to derive the proportion of ghost jobs, the distribution of search effort by age of ads, the distribution of ad renewal ages and the impact of obsolete information on unemployment and the efficiency of the matching process.	These rates will then be integrated into the structural models to infer the proportion of ghost posts, the age distribution of search efforts, the age distribution of ad renewals, and the impact of outdated information on unemployment and the effectiveness of the matching process.	These rates will then be integrated into structural models to deduce the proportion of ghost places, the distribution of age-searching ads, the distribution of ads' renewal ages and the impact of obsolete information on unemployment and the efficiency of the matching process.
	Evaluator 1		4 - Good	2 - Poor	3 - Fair
	Evaluator 2		3 - Fair	2 - Poor	2 - Poor
	Evaluator 3		4 - Good	2 - Poor	3 - Fair
	Evaluator 4		5 -Excellent	4 - Good	3 - Fair
40		Avec ses prouesses médiatisées, les techniques d'assistance médicale à la procréation (AMP) occupent la scène scientifique et politique.	With its media-friendly prowess, medically assisted reproductive techniques (MPA) are occupying the scientific and political stage.	With its high-profile prowess, assisted reproductive technology (art) occupies the scientific and political scene.	With its mediatized prowess, Assisted Reproductive Technologies (ART) occupy the scientific and political scene.

85	Evaluator 1		4 - Good	4 - Good	4 - Good
	Evaluator 2		2 - Poor	4 - Good	3 - Fair
	Evaluator 3		4 - Good	4 - Good	4 - Good
	Evaluator 4		4 - Good	3 - Fair	3 - Fair
	Evaluator 4		5 -Excellent	3 - Fair	4 - Good
281		[11] 139,6 millions de reais en 2006 (source : Relatorio administrativo da Escelsa, 2006) [12] à la date des entretiens, c'est-à-dire fin 2008.	[11] 139.6 million reais in 2006 (source: Relatorio administrativo da Escelsa, 2006) [12] at the time of the interviews, i.e. end of 2008.	[11] R $139.6 million in 2006 (source: Relatorio administrativo da Escelsa, 2006) [12] on the date of the interviews, i.e. at the end of 2008.	11 million reais [139,6] in 2006 (source: Relatorio administrativo da Escelsa, 2006) [12] on the date of interviews, i.e. at the end of 2008.
	Evaluator 1		4 - Good	3 - Fair	2 - Poor
	Evaluator 2		4 - Good	4 - Good	1 - Very Poor
	Evaluator 3		4 - Good	3 - Fair	2 - Poor
	Evaluator 4		5 -Excellent	5 - Excellent	1 - Very Poor
286		L'électroménager, neuf, est plus efficient et consomme moins d'énergie.	New appliances are more efficient and consume less energy.	The new home appliance is more efficient and consumes less energy.	The electrical appliance, nine, is more efficient and consumes less energy.
	Evaluator 1		5 -Excellent	4 - Good	2 - Poor
	Evaluator 2		5 -Excellent	2 - Poor	1 - Very Poor
	Evaluator 3		5 -Excellent	4 - Good	2 - Poor
	Evaluator 4		5 -Excellent	3 - Fair	1 - Very Poor
290		Les exploitants à proximité de Lima se retrouvent dans les années 1980 à la tête de parcelles comprises entre 3 et 5 ha (Mesclier, 2000).	In the 1980s, farmers near Lima found themselves in charge of plots of between 3 and 5 ha (Mesclier, 2000).	Farmers near Lima are found in the 1980s at the head of plots between 3 and 5 ha (Mesclier, 2000).	Farmers near Lima end in the 1980's with the head of plots between 3 and 5 ha (Mesclier, 2000).
	Evaluator 1		5 -Excellent	3 - Fair	2 - Poor
	Evaluator 2		4 - Good	3 - Fair	2 - Poor
	Evaluator 3		5 -Excellent	3 - Fair	2 - Poor
	Evaluator 4		5 -Excellent	3 - Fair	2 - Poor
302		Lors de la première apparition de la carte des taux d'occupation des lits de réanimation par des patients COVID-19 le 19 avril (Figure 1), Édouard Philippe explique : « si on présente la situation aujourd'hui en termes d'occupation des lits de réanimation, nous avons cette carte, qui montre que la stratégie de confinement et donc de limitation de circulation du virus a correctement fonctionné, ce dont nous devons nous réjouir.	When the map of the occupancy rates of resuscitation beds by COVID-19 patients first appeared on 19 April (Figure 1), Édouard Philippe explained: "if we present the situation today in terms of occupancy of resuscitation beds, we have this map, which shows that the strategy of containment and therefore limitation of the circulation of the virus has worked properly, which we should be pleased about.	During the first appearance of the map of the rates of occupancy of resuscitation beds by COVID-19 patients on April 19 (Figure 1), Édouard Philippe explains: "If we present the situation today in terms of occupancy of resuscitation beds, we have this map, which shows that the strategy of containment and therefore limitation of circulation of the virus has worked correctly, which we must rejoice.	During the first appearance of the map of the occupancy rates of intensive beds by COVID-19 patients on April 19 (Figure 1), Édouard Philippe explains: "if we present the situation today in terms of occupation of the intensive beds, we have this map, which shows that the confinement strategy and thus limitation of circulation of the virus has correctly functioned, which we must rejoice.
	Evaluator 1		4 - Good	3 - Fair	4 - Good

	Evaluator 2		2 - Poor	2 - Poor	2 - Poor
	Evaluator 3		4 - Good	3 - Fair	4 - Good
	Evaluator 4		3 - Fair	4 - Good	4 - Good

Table 7 – Segments from adequacy task, differences between engines

Productivity

High MT/PE difference

Segment	Source	MT output	Evaluator 1	Evaluator 2	Evaluator 3	Evaluator 4
12	En effet, le marché urbain via les circuits courts bénéficie avant tout aux producteurs organisés.	Indeed, the urban market via short circuits benefits above all organised producers. **= DeepL baseline**	Indeed, the urban market, via short circuits, primarily benefits organised producers. **MT/PE difference = 86.90%**	Indeed, the urban market, via local supply chains, benefits primarily organised producers. **MT/PE difference = 77.46%**	Indeed, the urban market via short circuits benefits above all organized producers. **MT/PE difference = 98.80%**	Indeed, the urban market through short circuits mainly benefits organised producers. **MT/PE difference = 83.83%**
83	Il sera fait une comparaison de l'énergie électrique dans deux quartiers distincts de la Région Métropolitaine du Grand Vitoria [1] .	It will be made a comparison of electrical energy in two distinct districts of the Metropolitan Region of the Grand Vitoria [1] . **= OpenNMT OPERAS + SciPar**	A comparison of electrical power supply in two separate districts of the Metropolitan Region of Greater Vitoria [1] will be made . **MT/PE difference = 74.9%**	Electrical energy supply will be compared in two separate districts of the Greater Vitoria Metropolitan Region [1] . **MT/PE difference = 56.33%**	It will be made a comparison of electrical energy in two distinct districts of the Metropolitan Region of the Grand Vitoria [1] . **MT/PE difference = 100%**	We will compare electrical energy in two districts of the Grand Vitoria Metropolitan Region [1] . **MT/PE difference = 65.49%**
133	Par exemple, le Monde, du 01/05/2020, https://www.lemonde.fr/planete/article/2020/05/01/des-erreurs-relevees-dans-la-premiere-cartographie-du-coronavirus_6038356_3244.html / article d'Europe 1, du 01/05/2020 : https://www.europe1.fr/sante/coronavirus-pourquoi-le-lot-apparait-il-en-rouge-sur-la-carte-du-deconfinement-3965581 / article de France Info, 01/05/2020 : https://www.franceinter.fr/societe/premiers-couacs-sur-la-carte-du-deconfinement-trois-departements-classes-rouges-a-cause-d-une-erreur?utm_medium=Social&utm_source=Facebook&fbclid=IwAR3nE95X2c3Kz-Jr2vWiuujkJcPaJ7Ppifth4e5yld9w_JTgaPZPAcp6EBg#Echobox=1588322963). [25] .	For example, Le Monde, du 01/05/2020, https://www.lemonde.fr/planete/article/2020/05/01/des-erreur-relevees-dans-la-premier-cartography-du-coronavirus_6038356_3244.html / article from Europe 1, of 01/05/2020 : https://www.europe1.fr/sante/coronavirus-pourquoi-le-porte-il-en **= OpenNMT OPERAS + SciPar**	For example, Le Monde, on 01/05/2020, https://www.lemonde.fr/planete/article/2020/05/01/des-erreur-relevees-dans-la-premiere-cartographie-du-coronavirus_6038356_3244.html / Europe 1 article, 01/05/2020: https://www.europe1.fr/sante/coronavirus-pourquoi-le-porte-il-en-rouge-sur-la-carte-du-deconfinement-3965581 / France Info article, 01/05/2020: https://www.franceinter.fr/societe/premiers-couacs-sur-la-carte-du-deconfinement-trois-departements-classes-rouges-a-cause-d-une-erreur?utm_medium=Social&utm_source=Facebook&fbclid=IwAR3nE95X2c3Kz-Jr2vWiuujkJcPaJ7Ppifth4e5yld9w_JTgaPZPAcp6EBg#Echobox=1588322963). [25] . **MT/PE difference = 56.95%**	For example, Le Monde, 1 May 2020: https://www.lemonde.fr/planete/article/2020/05/01/des-erreur-relevees-dans-la-premier-cartography-du-coronavirus_6038356_3244.html; Europe 1, 1 May 2020: https://www.europe1.fr/sante/coronavirus-pourquoi-le-porte-il-en **MT/PE difference = 91.46%**	For example, Le Monde, May 1, 2020, https://www.lemonde.fr/planete/article/2020/05/01/des-erreur-relevees-dans-la-premier-cartography-du-coronavirus_6038356_3244.html / article from Europe 1, of 01/05/2020 : https://www.europe1.fr/sante/coronavirus-pourquoi-le-porte-il-en **MT/PE difference = 97.14%**	For example, Le Monde from 05/01/2020, https://www.lemonde.fr/planete/article/2020/05/01/des-erreur-relevees-dans-la-premier-cartography-du-coronavirus_6038356_3244.html / article from Europe 1, of 05/01/2020 : https://www.europe1.fr/sante/coronavirus-pourquoi-le-porte-il-en-rouge-sur-la-carte-du-deconfinement-3965581 / article from France Info, 05/01/2020 : https://www.franceinter.fr/societe/premiers-couacs-sur-la-carte-du-deconfinement-trois-departements-classes-rouges-a-cause-d-une-erreur?utm_medium=Social&utm_source=Facebook&fbclid=IwAR3nE95X2c3Kz-Jr2vWiuujkJcPaJ7Ppifth4e5yld9w_JTgaPZPAcp6EBg#Echobox=1588322963). [25] . **MT/PE difference = 58.88%**
245	Durant l'année académique 2013-2014, l'Université libre de Bruxelles, l'Université Saint-Louis	During the 2013-2014 academic year, the Université libre de Bruxelles, the Université Saint-Louis and the	During the 2013-2014 academic year, the Université libre de Bruxelles, the Université Saint-Louis and the	During the 2013-2014 academic year, the Université libre de Bruxelles, the Université Saint-	During the 2013-2014 academic year, the Université libre de Bruxelles, the Université Saint-	During the 2013-2014 academic year, the Université libre de Bruxelles, the Université Saint-Louis and the medical field of the

	et le domaine médical de l'Université catholique de Louvain (basé à Woluwe) accueillaient respectivement 5550, 250 et 1550 étudiants ressortissants de l'Union européenne hors Belges, c'est-à-dire 23 %, 9 % et 28 % des inscriptions (données CREF).	medical field of the Université catholique de Louvain (based in Woluwe) welcomed 5550, 250 and 1550 non-Belgian EU students respectively, i.e. 23%, 9% and 28% of enrolments (CREF data). **= DeepL baseline**	medical field of the Université catholique de Louvain (based in Woluwe) welcomed 5550, 250 and 1550 non-Belgian EU students respectively, i.e. 23%, 9% and 28% of enrolments (CREF data). **MT/PE difference = 40.98%**	Louis and the medical faculty at the Université catholique de Louvain (based in Woluwe) welcomed 5 550, 250 and 1 550 non-Belgian EU students, respectively, i.e. 23%, 9% and 28% of all enrollments (CREF data). **MT/PE difference = 96.66%**	Louis and the medical field of the Université catholique de Louvain (based in Woluwe) welcomed 5550, 250 and 1550 non-Belgian EU students respectively, i.e. 23%, 9% and 28% of enrolments (CREF data). **MT/PE difference = 100%**	Université catholique de Louvain (based in Woluwe) received 5550, 250 and 1550 non-Belgian EU students respectively, i.e. 23%, 9% and 28% of enrolments (CREF data). **MT/PE difference = 98.64%**
441	En 2010, les travailleuses des titres-services étaient 75 % d'étrangères à Bruxelles, avec les Polonaises comme première nationalité.	In 2010, women workers in service vouchers were 75% foreigners in Brussels, with Polish women as their first nationality. **= ModernMT fine-tuned**	In 2010, 75% of service voucher workers in Brussels were foreigners, with Polish as the main nationality. **MT/PE difference = 71.68%**	In 2010, 75% of female workers paid in service vouchers in Brussels were foreign nationals, with Polish as their first nationality. **MT/PE difference = 73.02%**	In 2010, women workers in service vouchers were 75% foreigners in Brussels, with Polish women as their first nationality. **MT/PE difference = 100%**	In 2010, 75% of women workers in service vouchers in Brussels were foreigners, mostly Polish women. **MT/PE difference = 65.16%**

Table 8 – Segments from post-editing task, high MT/PE difference

Annex VIII: Self-paced reading experiment

Cumulative presentation of the text to the participants:

```
The increase in wild boar densities has a considerable economic cost and leads to a
 strong social tension between the stakeholders.
```

```
The increase in wild boar densities has a considerable economic cost and leads to a
 strong social tension between the stakeholders.
This tension, along with the lack of factual data, limits the possibility of
 collectively imagining other management practices.
```

```
The increase in wild boar densities has a considerable economic cost and leads to a
 strong social tension between the stakeholders.
This tension, along with the lack of factual data, limits the possibility of
 collectively imagining other management practices.
The project aims to cooperate stakeholders based on a study of the spatial ecology of
 the wild boar offering data playing the role of 'border objects' around which to
 articulate the discussion.
```

```
The increase in wild boar densities has a considerable economic cost and leads to a
 strong social tension between the stakeholders.
This tension, along with the lack of factual data, limits the possibility of
 collectively imagining other management practices.
The project aims to cooperate stakeholders based on a study of the spatial ecology of
 the wild boar offering data playing the role of 'border objects' around which to
 articulate the discussion.
A participative modeling will complete the approach, allowing to collectively build a
 vision of the system operation and explore management methods.
```

```
The increase in wild boar densities has a considerable economic cost and leads to a
 strong social tension between the stakeholders.
This tension, along with the lack of factual data, limits the possibility of
 collectively imagining other management practices.
The project aims to cooperate stakeholders based on a study of the spatial ecology of
 the wild boar offering data playing the role of 'border objects' around which to
 articulate the discussion.
A participative modeling will complete the approach, allowing to collectively build a
 vision of the system operation and explore management methods.
By revealing the socio-ecological interdependencies, the project will improve the
 synergy between actors in the co-construction and the implementation of more efficient
 management practices, and could be a significant advance in animal geography and spatial
 ecology, and in return for the management of human-wildlife conflicts.
```

```
The increase in wild boar densities has a considerable economic cost and leads to a
 strong social tension between the stakeholders.
This tension, along with the lack of factual data, limits the possibility of
 collectively imagining other management practices.
The project aims to cooperate stakeholders based on a study of the spatial ecology of
 the wild boar offering data playing the role of 'border objects' around which to
 articulate the discussion.
A participative modeling will complete the approach, allowing to collectively build a
 vision of the system operation and explore management methods.
By revealing the socio-ecological interdependencies, the project will improve the
 synergy between actors in the co-construction and the implementation of more efficient
 management practices, and could be a significant advance in animal geography and spatial
 ecology, and in return for the management of human-wildlife conflicts.
[END_OF_TEXT]
```

Figure 29 - Cumulative presentation of the text to the participants

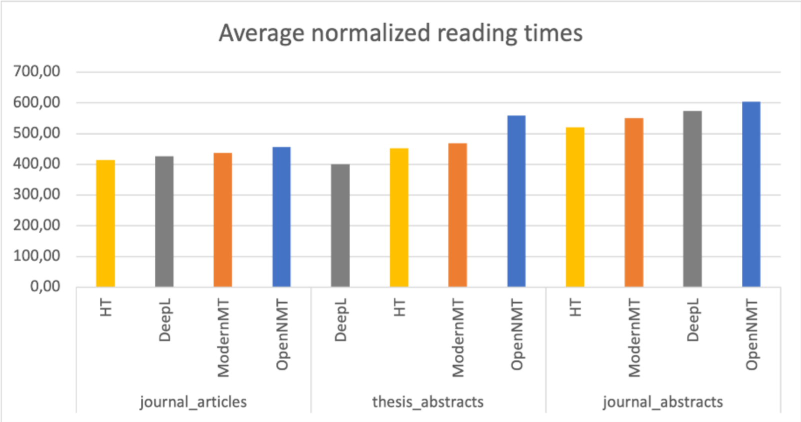


Figure 30 - Average normalized reading times

Annex IX: MQM error annotation process

The same data set that has been used for the self-paced reading experiments was manually analyzed for machine translation errors using the annotation platform Label Studio. This platform was locally installed on the UGent servers. The input format and taxonomy were configured:

Figure 31 - Input format and taxonomy in Label Studio

The annotation guidelines were then prepared, followed by a meeting with the evaluator and tests:

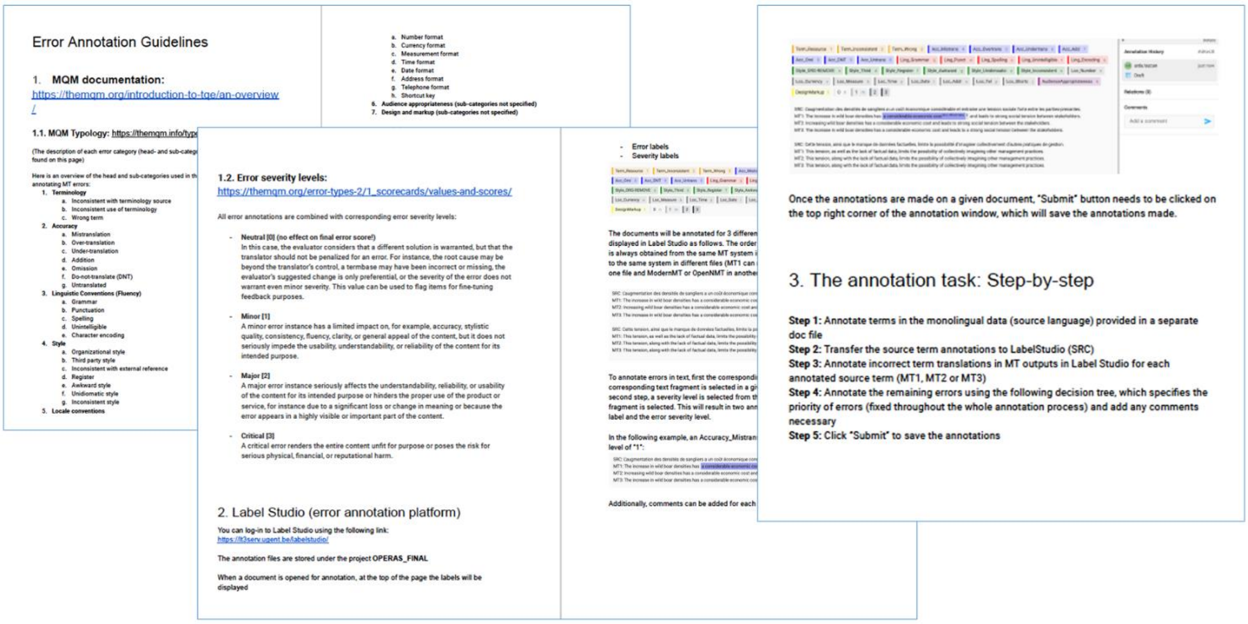
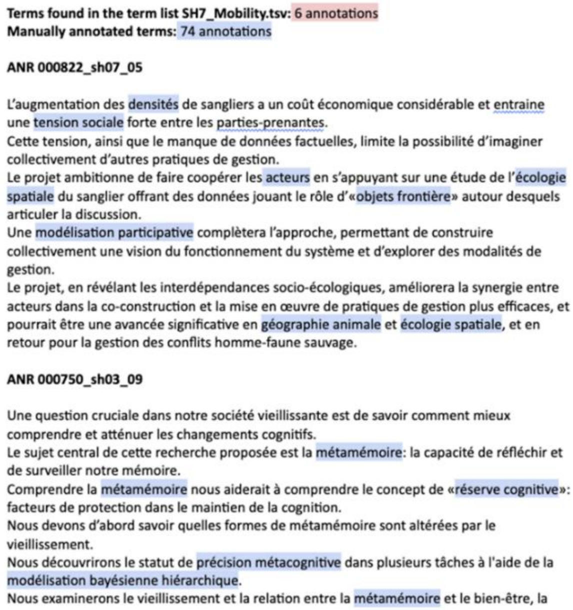


Figure 32 - Annotation guidelines for MQM

Prior to error annotation, terms were marked in the source texts (a) automatically using a Python script (based on the term lists provided for the discipline in question, the terms being lemmatized and lowercased except in case of fully uppercase entries), and (b) by the annotator using the term extraction methodology proposed by Rigouts Terryn et al. (2020) which was discussed in deliverable D1. Figure 33 illustrates the annotation results. The number of terms marked during both steps are as follows:

- Terms marked using the term list SH7_Mobility.tsv: 6
- Terms marked by the annotator: 74



Terms found in the term list SH7_Mobility.tsv: 6 annotations
Manually annotated terms: 74 annotations

ANR 000822_sh07_05

L’augmentation des densités de sangliers a un coût économique considérable et entraine une tension sociale forte entre les parties-prenantes.
Cette tension, ainsi que le manque de données factuelles, limite la possibilité d’imaginer collectivement d’autres pratiques de gestion.
Le projet ambitionne de faire coopérer les acteurs en s’appuyant sur une étude de l’écologie spatiale du sanglier offrant des données jouant le rôle d’«objets frontière» autour desquels articuler la discussion.
Une modélisation participative complètera l’approche, permettant de construire collectivement une vision du fonctionnement du système et d’explorer des modalités de gestion.
Le projet, en révélant les interdépendances socio-écologiques, améliorera la synergie entre acteurs dans la co-construction et la mise en œuvre de pratiques de gestion plus efficaces, et pourrait être une avancée significative en géographie animale et écologie spatiale, et en retour pour la gestion des conflits homme-faune sauvage.

ANR 000750_sh03_09

Une question cruciale dans notre société vieillissante est de savoir comment mieux comprendre et atténuer les changements cognitifs.
Le sujet central de cette recherche proposée est la métamémoire: la capacité de réfléchir et de surveiller notre mémoire.
Comprendre la métamémoire nous aiderait à comprendre le concept de «réserve cognitive»: facteurs de protection dans le maintien de la cognition.
Nous devons d’abord savoir quelles formes de métamémoire sont altérées par le vieillissement.
Nous découvrirons le statut de précision métacognitive dans plusieurs tâches à l'aide de la modélisation bayésienne hiérarchique.
Nous examinerons le vieillissement et la relation entre la métamémoire et le bien-être, la

Figure 33 - Annotated term examples

Subsequently, the Label Studio files were prepared using a Python script. The MT order was randomised for each file. This is illustrated in Figure 34.

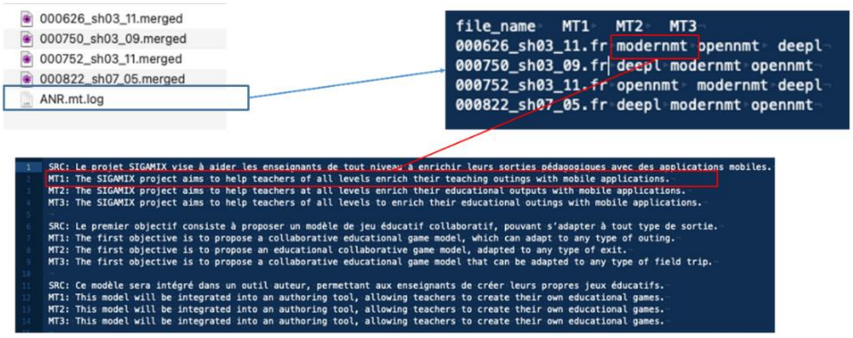


Figure 34 - Preparation of the Label Studio files using Python

The term annotations discussed above were transferred to Label Studio, leading term errors to be annotated on this platform (1st priority in MQM decision tree). The annotator then labelled other types of errors, as shown in Figure 35.

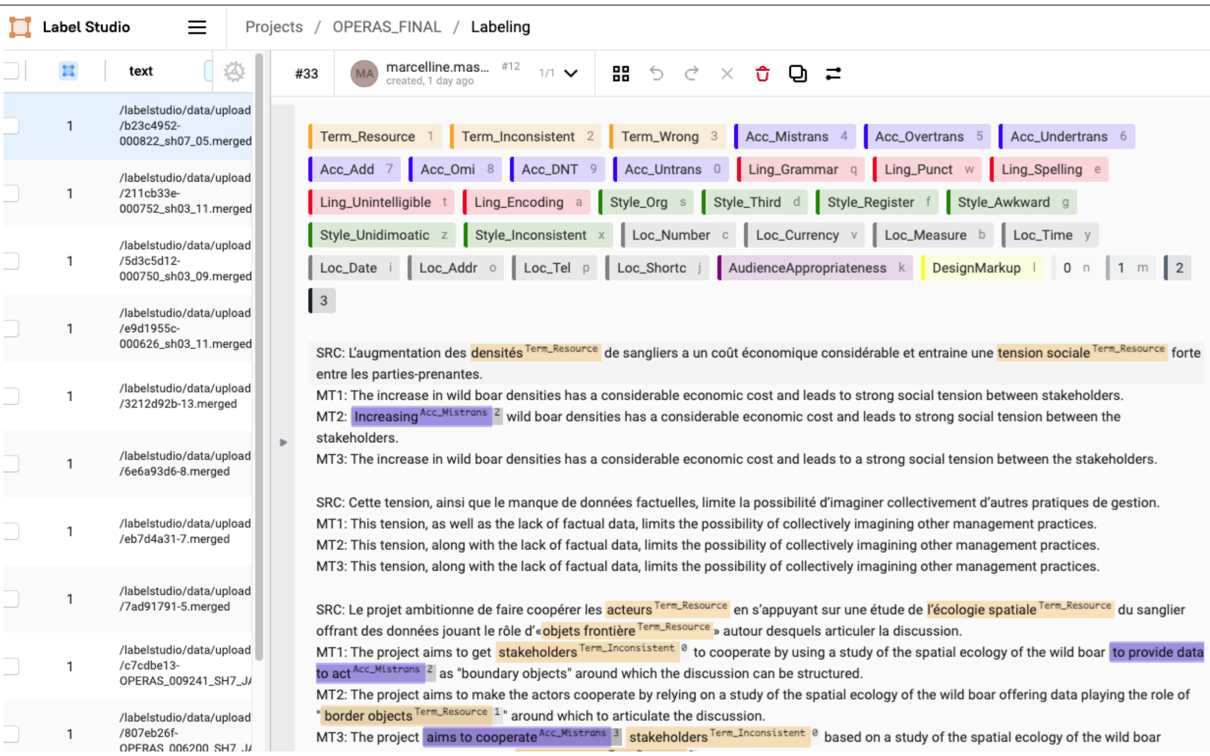


Figure 35 - Types of annotated errors

Subsequently, the error annotations were exported to JSON:

Figure 36 - Error annotations JSON export

The results were analyzed per text type and for the whole evaluation set, using Python and Excel. These results can be presented in two categories: (i) MQM scorecards, and (ii) other analyses. In Figure 37, we show the conversion of the JSON data to severity counts, which were then further used for producing the scorecards and the other analyses.

Figure 37 - Conversion of the JSON data to severity counts

Annex X: MQM error annotation results

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

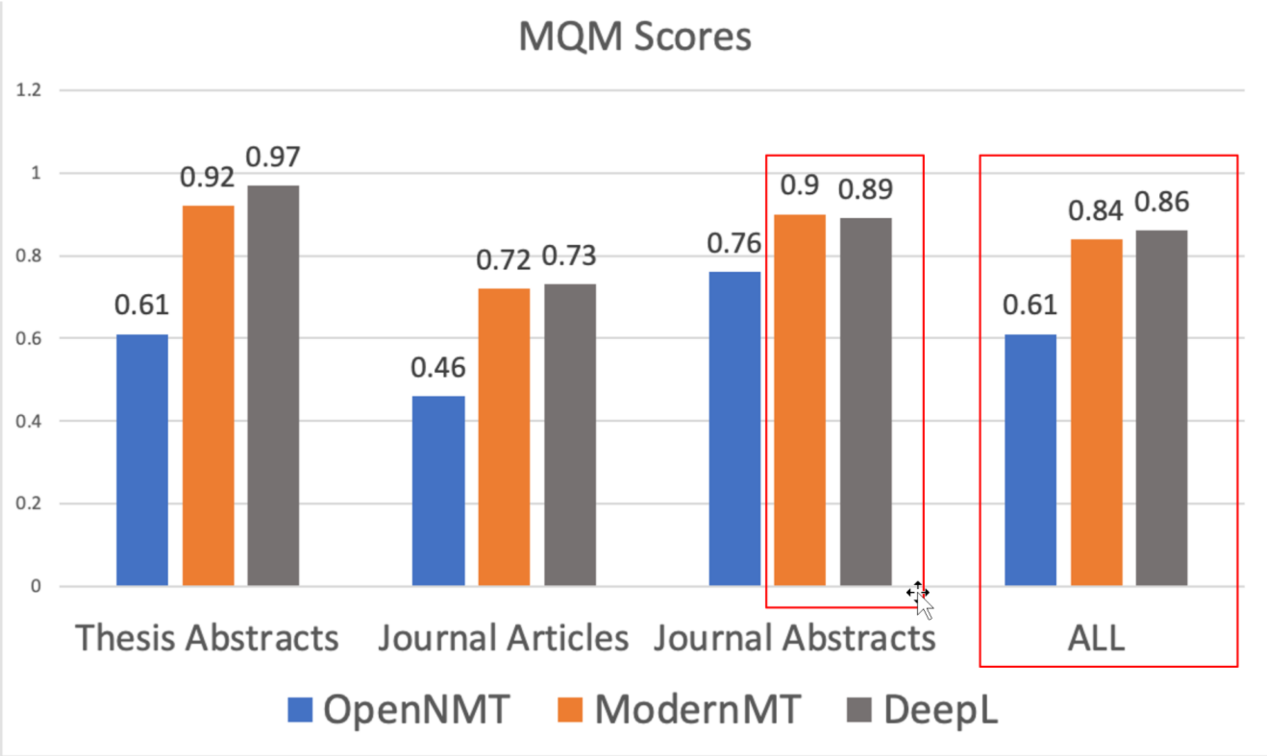
Domain	MT System					
ALL	OpenNMT					
Error Severity Levels:	**Neutral**	**Minor**	**Major**	**Critical**	**Error Type Penalty Total**	
Severity Multipliers:	0	1	5	25		
Error Types	**Error Counts**				ET Weights	**ETPTs**
Term_Resource	7	**6**	0	3	1	81.0
Term_Inconsistent	**2**	0	0	0	1	0.0
Term_Wrong	0	1	1	1	1	31.0
Acc_Mistrans	2	8	**5**	**18**	1	483.0
Acc_Overtrans	0	0	**1**	1	1	30.0
Acc_Undertrans	0	**3**	0	**1**	1	28.0
Acc_Add	0	1	1	0	1	6.0
Acc_Omi	0	4	0	0	1	4.0
Acc_DNT	0	0	0	0	1	0.0
Acc_Untrans	0	0	0	0	1	0.0
Ling_Grammar	0	3	1	0	1	8.0
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	1	**1**	0	1	6.0
Style_Awkward	1	3	4	0	1	23.0
Style_Unidimoatic	0	3	3	0	1	18.0
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	2	0	1	10.0
Loc_Date	0	0	1	0	1	5.0
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):			738.00
Evaluation Word Count (EWC):	**1892**		**Per-Word Penalty Total (PWPT):**			0.3901
Reference Word Count (RWC):	1000		**Overall Normed Penalty Total (ONPT):**			390.06
Penalty Scaler (PS):	1.00		**Overall Quality Score (OQS):**			60.99
Max. Score Value (MSV):	100.00					
				Overall Quality Fraction		0.61
Total no. of errors	90			Sentences with errors		57.00
Total critical errors	24			Total sentences		83.00
				% Sentences with errors		0.69

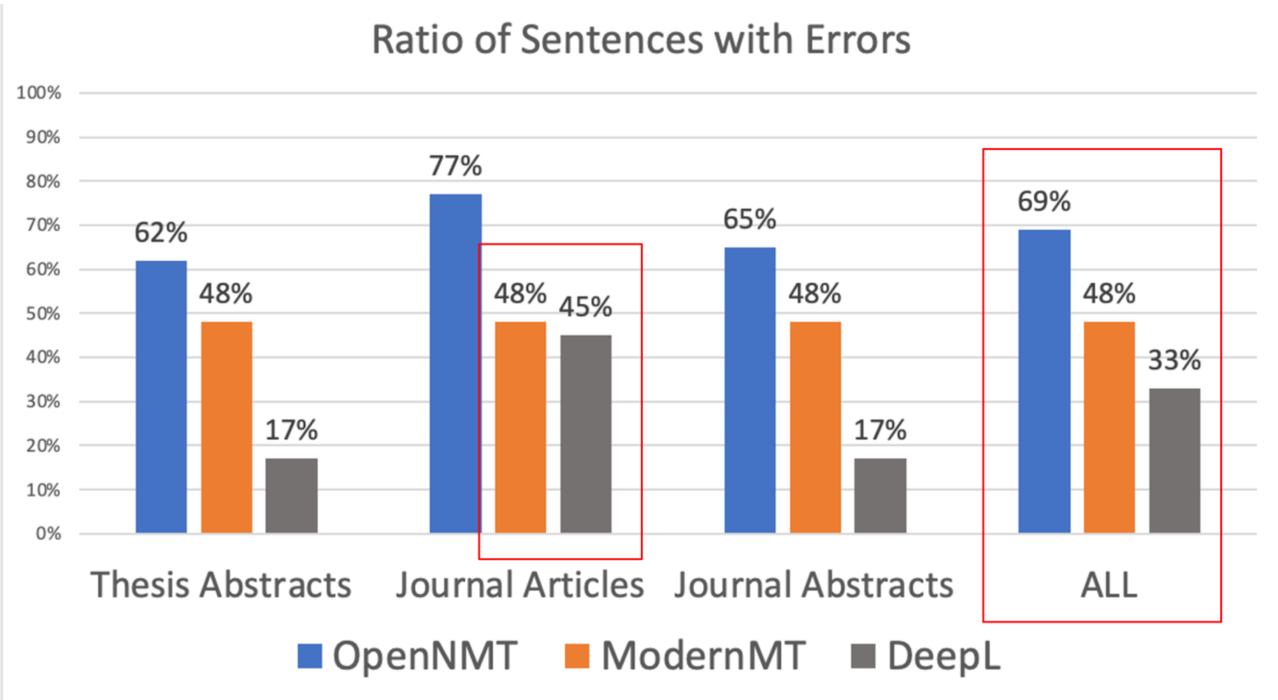
Domain	MT System					
ALL	ModernMT					
Error Severity Levels:	**Neutral**	**Minor**	**Major**	**Critical**	**Error Type Penalty Total**	
Severity Multipliers:	0	1	5	25		
Error Types	**Error Counts**				ET Weights	**ETPTs**
Term_Resource	3	6	0	1	1	31.0
Term_Inconsistent	1	0	0	0	1	0.0
Term_Wrong	0	0	0	0	1	0.0
Acc_Mistrans	0	2	7	7	1	212.0
Acc_Overtrans	1	0	0	1	1	25.0
Acc_Undertrans	0	4	0	0	1	4.0
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	2	1	0	1	7.0
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	3	0	0	1	3.0
Style_Awkward	0	4	3	0	1	19.0
Style_Unidimoatic	0	2	0	0	1	2.0
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):			308.00
Evaluation Word Count (EWC):	1909		**Per-Word Penalty Total (PWPT):**			0.1613
Reference Word Count (RWC):	1000		**Overall Normed Penalty Total (ONPT):**			161.34
Penalty Scaler (PS):	1.00		**Overall Quality Score (OQS):**			83.87
Max. Score Value (MSV):	100.00					
				Overall Quality Fraction		0.84
Total no. of errors	49			Sentences with errors		40.00
Total critical errors	9			Total sentences		83.00
				% Sentences with errors		0.48

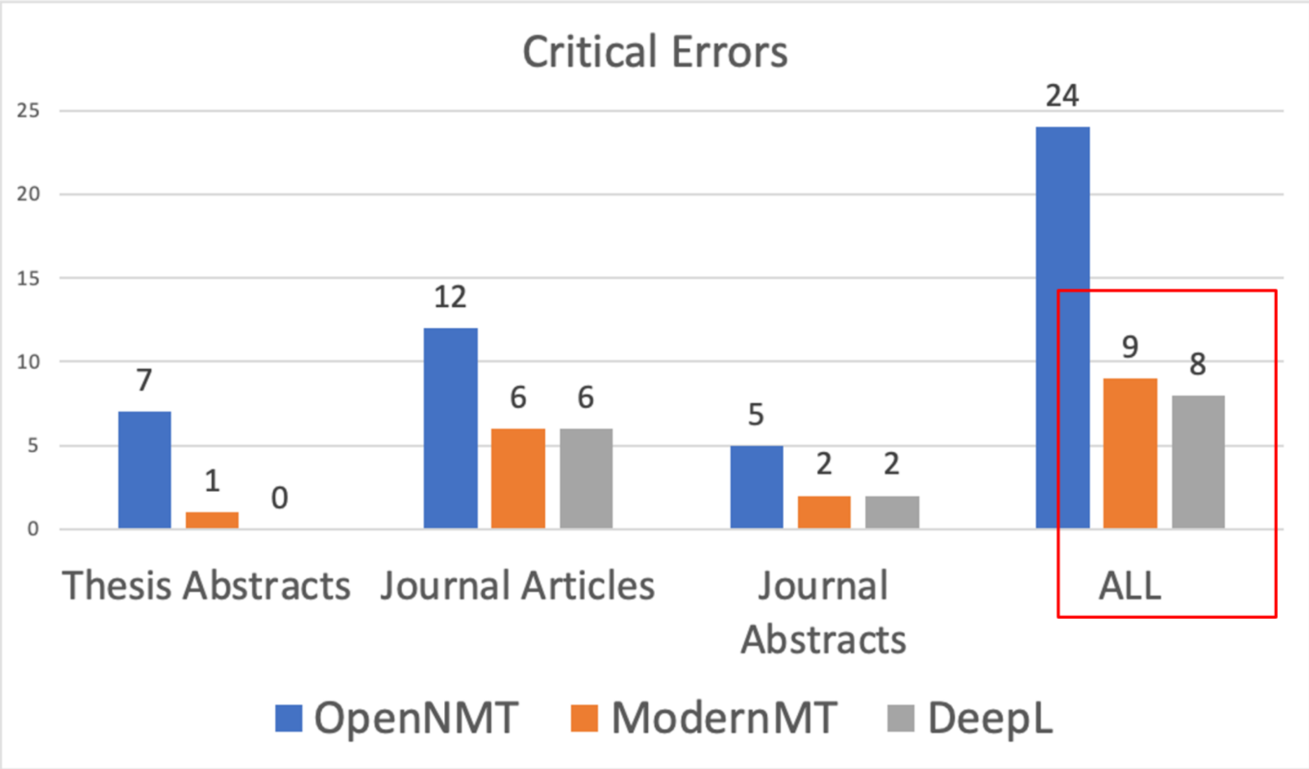
Domain	MT System					
ALL	DeepL					
Error Severity Levels:	**Neutral**	**Minor**	**Major**	**Critical**	**Error Type Penalty Total**	
Severity Multipliers:	0	1	5	25		
Error Types	**Error Counts**				ET Weights	**ETPTs**
Term_Resource	4	3	0	2	1	53.0
Term_Inconsistent	1	1	0	0	1	1.0
Term_Wrong	0	0	0	0	1	0.0
Acc_Mistrans	0	2	**4**	5	1	147.0
Acc_Overtrans	0	0	**1**	1	1	30.0
Acc_Undertrans	0	3	0	0	1	3.0
Acc_Add	0	0	1	0	1	5.0
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	0	2	0	1	10.0
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	2	2	0	1	12.0
Style_Unidimoatic	0	1	2	0	1	11.0
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):			277.00
Evaluation Word Count (EWC):	1922		**Per-Word Penalty Total (PWPT):**			0.1441
Reference Word Count (RWC):	1000		**Overall Normed Penalty Total (ONPT):**			144.12
Penalty Scaler (PS):	1.00		**Overall Quality Score (OQS):**			85.59
Max. Score Value (MSV):	100.00					
				Overall Quality Fraction		0.86
Total no. of errors	38			Sentences with errors		27.00
Total critical errors	8			Total sentences		83.00
				% Sentences with errors		0.33

Table 9 - MQM scorecards regarding all evaluation data, per MT engine

The results of other analyses are provided per text type and for the whole evaluation set, per MT engine, see Figure 38 (the information in the two first graphs also appears in the MQM scorecards).


MQM Scores
1.2
1
0.8
0.6
0.4
0.2
0
0.61
0.92
0.97
0.46
0.72
0.73
0.76
0.9
0.89
0.61
0.84
0.86
Thesis Abstracts
Journal Articles
Journal Abstracts
ALL
OpenNMT
ModernMT
DeepL


Ratio of Sentences with Errors
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
62%
48%
17%
77%
48%
45%
65%
48%
17%
69%
48%
33%
Thesis Abstracts
Journal Articles
Journal Abstracts
ALL
OpenNMT
ModernMT
DeepL



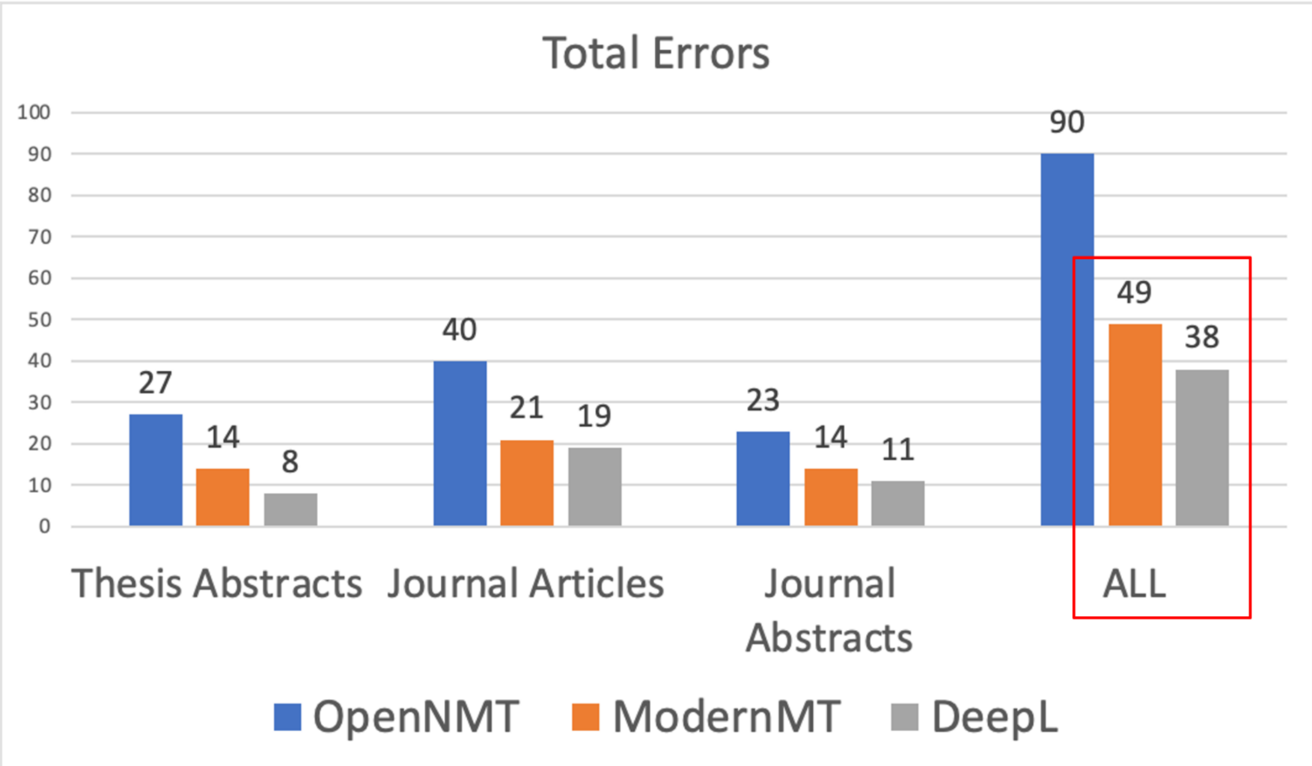


Figure 38 - MQM annotation results per text type and for the whole evaluation set, per MT engine

CROSSLANG

www.crosslang.com

CrossLang NV
Amerikagebouw Kerkstraat
106 9050 Gentbrugge
Belgium

+ 32 9 335 22 00
info@crosslang.com