Translations and Open Science

Study on machine translation evaluation
in the context of scholarly communication

D4: Outcome for discipline

"Climatology and climate change"

Version: final

Authors:

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

Lieve Macken, Arda Tezcan (LT3)

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

CROSSLANG

language and translation technology team

DISCLAIMER

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

AVERTISSEMENT

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

Table of contents

1. Introduction

This deliverable outlines the evaluation outcome and best practices for the discipline “Climatology and climate change” (ERC code PE10). In particular, it provides the following description of the data, models and results obtained for this discipline using the procedure outlined in deliverable D1: statistics regarding the training and test material selected for this discipline, information concerning the engines trained, and scores produced using automated MT metrics.

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

2. Training and fine-tuning MT engines

2.1. Training and evaluation data

The data selected in call 1 consists of seven publication types from 388 different sources of publication, and a terminology list. Table 1 gives an overview of the size and distribution.

Type of publication	Documents	Segments
Book	1	1908
Conference paper abstract	134	739
Journal article[1]	103	645
Journal article abstract	1621	38177
Publication type not available	61	505
Report	6	14009
Thesis abstract	3703	44580
Terminology	-	397
Total	**5629**	**100960**

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 56 additional abstracts obtained from the **ANR dataset**[2] falling under discipline PE10 (*Climatology and climate change*). No additional popularizing articles were selected for PE10, unlike in case of discipline 1 and 2; a search by OPERAS for relevant bilingual material did not yield results due to the high sparseness of such material.

2.2. Data partitioning

The dataset for PE10 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] It should be noted that data from call 1 labeled as "journal article" for this discipline appear to have been labeled incorrectly, as they are actually journal article abstracts.

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

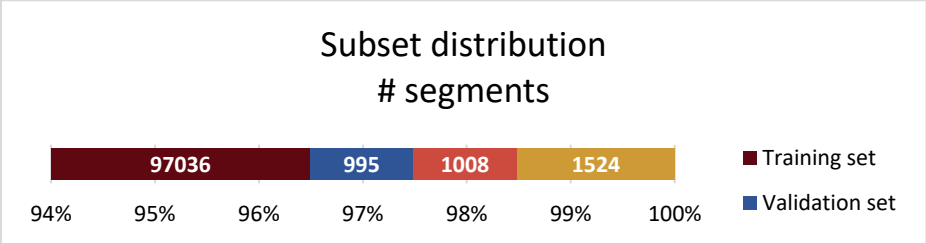


Figure 1 - Distribution of training, validation, testing, and evaluation sets

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

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

Validation set: It consists entirely of data from call 1. As we want a significant representation of each text type, 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: The same criteria as for the validation set apply.

Evaluation set: See Section 2.1.

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

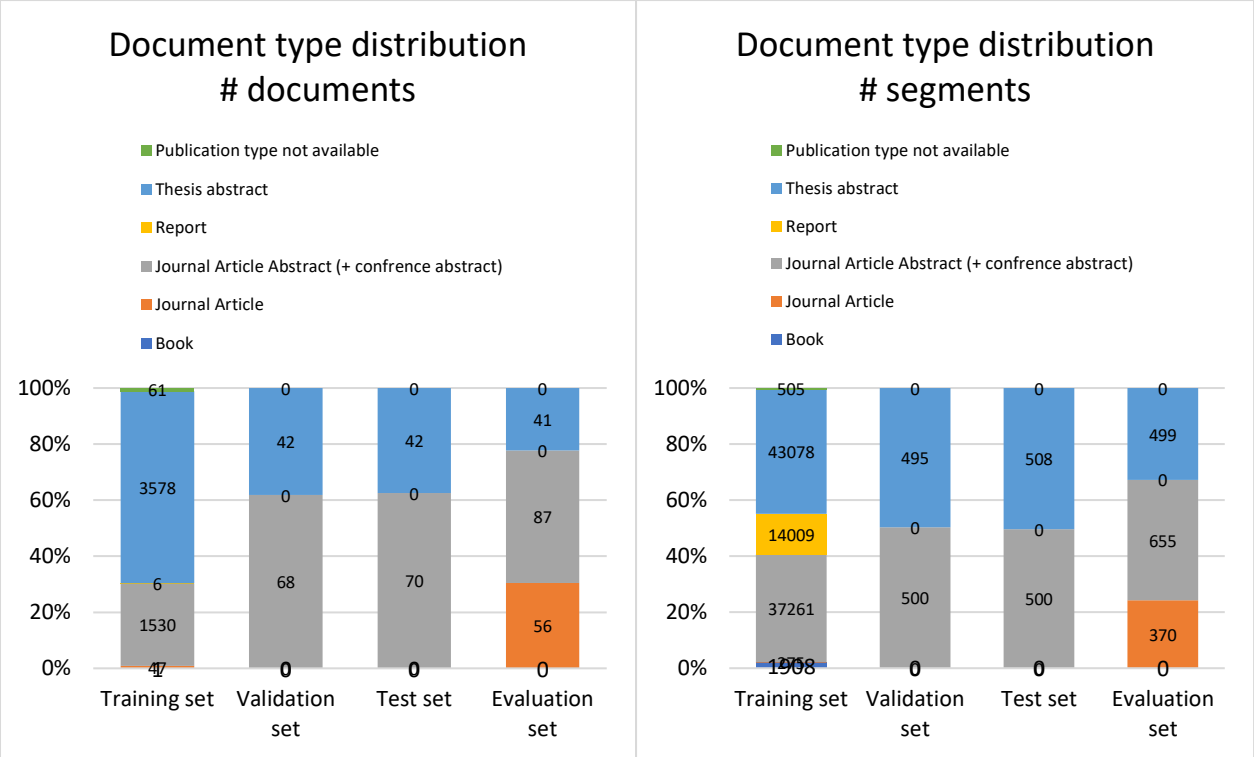


Figure 2 – Distribution of publication types for each subset, number of documents and segments

2.3. MT Customisations

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

Type	System	Short description[3]	Duration[4]	Date
commercial	DeepL	Baseline	/	19/04/2023
		custom (termbase)	5 seconds	19/04/2023
	ModernMT	Baseline	/	19/04/2023
		custom (OPERAS training data)	1 minute 30 seconds	19/04/2023
open source	OpenNMT	Baseline	3 h 20 m/iteration	03/05/2023
		custom 1 (OPERAS training data)	3 h 20 m/iteration	21/04/2023
		custom 2 (OPERAS training data + SciPar)	3 h 20 m/iteration	21/04/2023

Table 2 - Overview of the MT experiments

[3] 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.

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

3. Automated evaluation

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

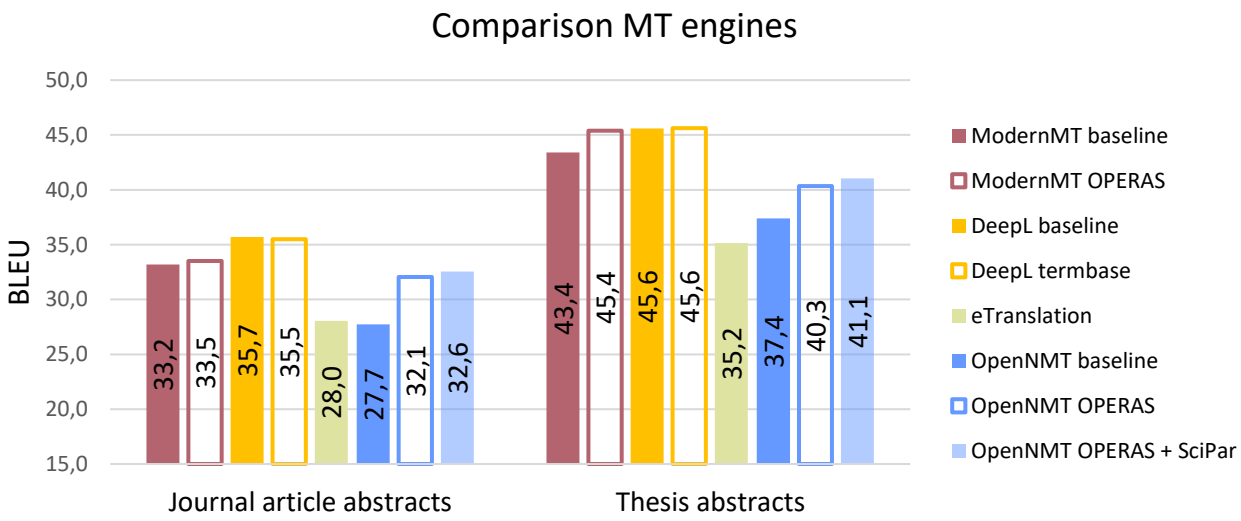


Figure 3 – Comparison of MT engines, using BLEU score, for each text type

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

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

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

4. Human evaluation

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

4.1. Setup and execution of adequacy task

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

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

4.2. Results of adequacy task

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

User ratings

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

Number of times each engine is ranked first

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

Correlations

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

4.3. Post-editing task

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

Post-editing times

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

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

When we look at the post-editing times per document type, we see that there is hardly any difference between the two types of abstracts.

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. There is little difference between the two types of abstracts.

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

HTER

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

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

4.4. Self-paced reading experiment

Twelve texts were selected for the discipline from three different sources (see Table 3). It was rather difficult to select suitable texts for lay persons as the discipline contained rather technical texts. Moreover, the human reference translations were often too divergent from the source or incorrect. Therefore, we were forced to manually correct the reference translation in one text (000003_pe10_03; changes were made to 3 segments). There were no full texts available for D3. The three sets consist of three different sources of abstracts.

CLIMATOLOGY	No. src words	No. segments
ANR ABSTRACTS		
000003_pe10_03	146	5
000819_pe10_03	146	5
000815_pe10_03	113	5
000824_pe10_03	131	6
TAUS journal abstracts		
OPERAS_000628	109	4
OPERAS_000281	105	5
OPERAS_000050	125	5
OPERAS_000017	108	4
TAUS THESIS ABSTRACTS		
OPERAS_000566_PE10_TA	121	6
OPERAS_000522_PE10_TA	140	6
OPERAS_000677_PE10_TA	137	5
OPERAS_005209_PE10_TA	129	7
TOTAL	**1510**	**64**

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

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

The experiments were carried out from June 1st until June 19th, 2023 (during the same time span, the experiments for discipline 2, see D3, were also carried out). The duration of the sessions varied between 60 and 80 minutes.

Translation quality was assessed as sufficient in 69% of all assessments (which is lower than discipline 1, i.e. Human Mobility, where the result was 74%). In 44 of the 144 assessments, translation quality was rated as insufficient, as indicated in Table 4.

Quality Score: no		Total
	DeepL	9
	ModernMT	11
	HT	12
	OpenNMT	12
Total		**44**

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

Average normalized reading times (ms per word) were highest for HT (665 ms). Within the MT systems, the average reading times were on par for OpenMT (597 ms) and DeepL (597 ms), followed by ModernMT (615 ms), although there is some variation across text types (see Figure 4).

Observed average normalized reading times were the highest for discipline 3 (compared to discipline 1 and 2). This can partially be explained by the text types (only abstracts, no full articles).

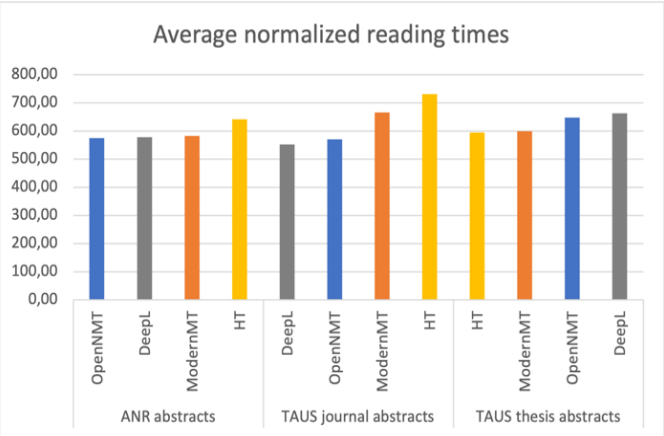


Figure 4 - Average normalized reading times

4.5. MQM error annotation

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

- Terms marked using the term list PE10_Climatology.tsv: 5
- Terms marked by the annotator: 116

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

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

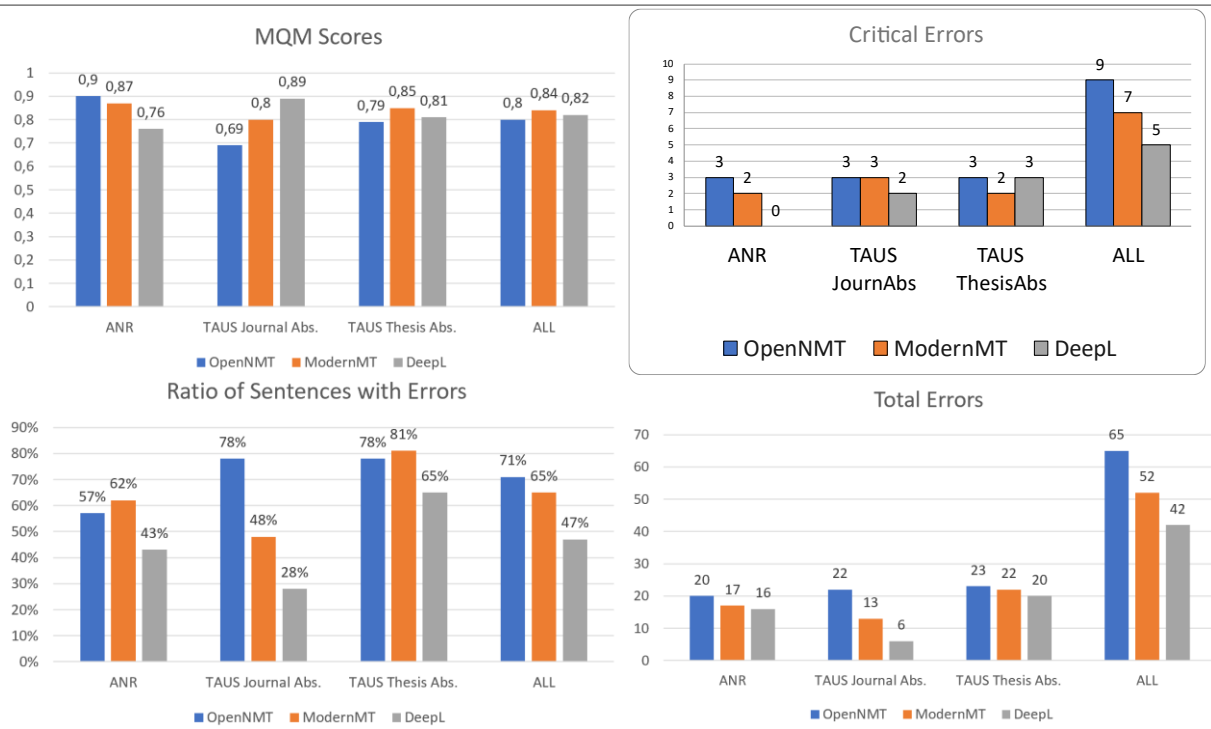


Figure 5 – Various types of scores resulting from manual error annotation

From the scorecards and analyses, we can conclude that we obtain a similar ranking for engines as for automatic evaluation scores in case of total errors and ratio of sentences with errors, i.e. DeepL scores better than ModernMT and OpenNMT.

5. Conclusions

In this deliverable, we presented detailed information on the third discipline, “Climatology and climate change”, 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 larger for ModernMT baseline and fine-tuned. The most significant difference was observed for OpenNMT fine-tuned (with and without SciPar data) and baseline. Overall, the scores for DeepL were the highest. In addition to the automatic scores, human evaluations were performed. Four types of tasks were performed in order to obtain the results (adequacy task, productivity task, self-paced reading experiment and MQM error annotation).

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

Results from the productivity task indicate that DeepL produces the best outputs. However, in terms of post-editing time, there is no significant difference between the engines. Translators take on average much longer to correct the text than the researchers. There is hardly any difference between the two types of abstracts involved (journal article abstracts and thesis abstracts) in terms of post-editing time. Furthermore, post-editing DeepL outputs showed the lowest average perceived effort, followed by ModernMT and OpenNMT. A correlation was observed between perceived effort and post-editing time, between perceived effort and HTER, and between HTER and post-editing time.

Regarding the self-paced reading experiment, it was rather difficult to select suitable texts for lay persons as the discipline contained rather technical texts. Translation quality was assessed as sufficient in 69% of all assessments. Average normalized reading times (ms per word) were highest for HT (665 ms). Within the MT systems, the average reading times were on par for OpenMT (597 ms) and DeepL (597 ms), followed by ModernMT (615 ms).

From the MQM scorecards and analyses, we can conclude that we obtain a similar ranking for engines as for automatic evaluation scores, i.e. DeepL scores better than ModernMT and OpenNMT.

Annex I: Dataset challenges and examples

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

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.

Thesis abstracts:

Source EN	Reference FR
3. How are snow physical characteristics related to the local SMB? 4. How can remote sensing data be best used to infer snow physical characteristics and SMB? 5. What are recent and future SMB distributions over Antarctica?	3. Les modèles de circulation atmosphériques sont-ils capables de reproduire les processus et la distribution du BMS? 3. Peut-on relier les caractéristiques physiques des couches de neige proche de la surface avec les variations locales du BMS?
The ASUMA project proposes to combine numerous field data and original techniques to better constrain remote sensing data and modeling results in the transition zone from the coast to the Antarctic plateau.	4. Peut-on retrouver caractériser l'état du manteau neigeux à partir des données de télédétection, et par suite, offrir des données pour interpoler les valeurs de BMS disponibles? Ce projet combine différentes approches.

Source EN	Reference FR
In this project, we propose an alternative method for quantifying the recent variations of SMB and their relationship with climate, atmospheric circulation and moisture origin. We will perform original field measurements of SMB and snow physics and robustly link them to satellite data. We will combine this information with the use of back-trajectories and regional to global modeling.	Nous proposons ici de mieux calibrer ces relations grâce à une méthode alternative basée sur une large palette de mesures originales effectuées dans la zone de transition, combinées à des analyses de rétro-trajectoires et de modélisation globale et régionale pour évaluer les variations récentes et future de BMS et les relier au climat.

Source EN	Reference FR
Clouds play a crucial role to regulate the surface energy budget with competing warming and cooling effects.	Les nuages jouent un rôle primordial sur le bilan radiatif en Arctique.

Journal article abstracts:

Source EN	Reference FR
Cropping practices and some soil characteristic amendments are suggested herein for this purpose.	L'auteur termine par quelques suggestions de pratiques culturales et de modifications des caractères du sol susceptibles d'atténuer les émissions.

Source EN	Reference FR
Our results reveal five different types of oil palm agroforestry systems: i) associations with livestock during the production phase of the oil palm; ii) traditional African palms and food crops systems sustained over time; iii) associations with food crops during the juvenile phase of the oil palm; iv) systems developed by family farms that permanently associate other plants; and v) prototype designs developed by research institutions, often at the request of local agricultural enterprises.	iii) l'agroforesterie temporaire avec des cultures vivrières en palmeraie juvénile ; iv) l'agroforesterie permanente avec des cultures pérennes ; et enfin v) des prototypes de systèmes agroforestiers à base de palmiers sélectionnés, conçus par des institutions de recherche et développement, souvent à la demande d'entreprises agricoles ou d'agro-industries.

Table 5 - Bad reference translations and misalignments

Segmentation issues

- **Sentences glued**

Thesis abstracts:

Source EN	Reference FR
A sensitivity analysis served to determine principal variables affecting SEB estimations and make recommendations for further campaigns of **measure.Second**, the fractal geometry was used to analyse the complexity and heterogeneity of the spatial distribution in the case study of Est-Ensemble (eastern of Paris), and to develop a multiscale scenario of NBS deployment.	Une analyse de sensibilité a permis de déterminer les principales variables affectant les estimations de SEB et de formuler des recommandations pour d'autres campagnes de mesure.Dans un second temps, la complexité et l'hétérogénéité de l'organisation spatiale d'un territoire où les SFN seraient implantées, a été étudiée à travers la géométrie fractale.

Source EN	Reference FR
The use of RHIZOtest allowed us to describe the effects of several environmental parameters at exposure concentration close to the ones predicted in **soil.The** RHIZOtest can be used as a tool to deign NNs for a controlled phytoavailability or for predicting the transfer in trophic chain as a function of the soil.	L'utilisation du RHIZOtest nous a donc permis de cribler l'effet d'un grand nombre de facteurs environnementaux sur la phytodisponibilité des NMs à des concentrations proches de celles prédites dans les sols par la modélisation.

Table 6 - Glued sentences in the data

- **Words glued**

Source EN	Reference FR
Can we use these **observationsto** document the surface-air exchanges of these compounds, which represent an **importantpart** of the water and carbon biogeochemical cycles?	Peut-on en extrairedes informations sur les échanges entre la surface et l'atmosphère, qui constituent une part **crucialedes** cycles biogéochimiques du carbone et de l'eau ? L'interprétation paléoclimatique **desenregistrements** de carottes de glace au Groenland peut-elle bénéficier de ces observations ?

Source EN	Reference FR
This facilitates the interpretation of **ourobservations** in terms of large scale atmospheric transport **signals.To** better understand the observed variations, I related our data series with other **observationsoriginating** from different sites, and with outputs from different atmospheric models.	Cette caractéristique facilite l'attribution des variations observées à des changements **detransport** atmosphérique de grande échelle en lien avec des sources **distantes.Pour** comprendre les variations observées, je les ai mises en relation avec des **observationsissues** d'autres sites ainsi qu'avec des sorties de modèles atmosphériques de grande échelle.

Table 7 - Glued words in the data

Freely translated outputs

Source EN	Reference FR
The results shows micro-UHI inside of Paris from 2 to 4°C and the vulnerability of the peripheral districts of Paris and some surrounding cities due to urban plume (up to 2°C warmer).	Les résultats permettent de discerner des micro-ICU au sein de Paris avec des différences de température de 2 à 4°C et mettent en évidence la vulnérabilité des arrondissements périphériques et de certaines communes limitrophes en liaison avec le panache urbain engendrant des différences de température de l'ordre de 2°C.

Table 8 - Example of freely translated outputs

Evaluation Setup Challenges

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

Source EN	Reference FR	ModernMT
Using large ensembles of IPSLCM5A model simulations, we first investigate the roles of internal variability (and in particular the IPO) and external forcing in driving recent Peru-Chile regional **cooling.The simulations** reproduce the relative cooling, in response to an externally-forced southerly wind anomaly, which strengthens the upwelling off Chile in recent decades.	Nous avons d'abord étudié les rôles de la variabilité interne (telle que l'IPO) et des forçages externes dans le refroidissement régional dans la zone d'upwelling du Pérou-Chili et les mécanismes associés.	En utilisant de grands ensembles de simulations de modèles IPSLCM5A, nous étudions d'abord les rôles de la variabilité interne (et en particulier de l'IPO) et du forçage externe dans le refroidissement régional récent entre le Pérou et le Chili.

Source EN	Reference FR	DeepL
South Greenland is a key region placed under Arctic and Northern Atlantic influences, whichis poorly documented in terms of atmospheric **monitoring.The** aim of my thesis is to conductand use the first regional atmospheric observations of CO2, O2, CH4 and isotopic compositionof water vapour and precipitation performed in Ivittuut, a coastal site in south-west Greenland.	Le but de ma thèse est de conduire et d'utiliser les premières observations atmosphériquesde surface du CO2, de l'O2, du CH4 et de la composition isotopique de la vapeur d'eau et desprécipitations dans la région peu instrumentée du sud du Groenland.	Le but de ma thèse est de réaliser et d'utiliser les premières observations atmosphériques régionales de CO2, O2, CH4 et de la composition isotopique de la vapeur d'eau et des précipitations effectuées à Ivittuut, un site côtier du sud-ouest du Groenland.

Source EN	Reference FR	DeepL	OpenNMT
With the Arctic warming twice as fast as the globe, the present and future variability of snow characteristics are crucially important for better understanding of the processes and changes undergoing with **climate.However,** our capacity to observe the terrestrial Arctic is limited compared to the mid-latitudes and climate models play very important role in our ability to understand the snow-related processes especially in the context of a warming cryosphere.	Le réchauffement de l'Arctique étant deux fois plus rapide que celui du reste du globe, la variabilité présente et future des caractéristiques de la neige est cruciale pour une meilleure compréhension des processus et des changements climatiques.Cependant, notre capacité à observer l'Arctique terrestre étant limitée, les modèles climatiques jouent un rôle clé dans notre aptitude à comprendre les processus liés à la neige.	Cependant, notre capacité d'observation de l'Arctique terrestre est limitée par rapport aux latitudes moyennes et les modèles climatiques jouent un rôle très important dans notre capacité à comprendre les processus liés à la neige, en particulier dans le contexte d'une cryosphère qui se réchauffe.	Étant donné le réchauffement de l'Arctique deux fois plus rapide que le globe, la variabilité actuelle et future des caractéristiques de la neige est cruciale pour une meilleure compréhension des processus et des changements subis avec le climat.

Source EN	Reference FR	DeepL
In this respect representation of snow-associated feedbacks in climate models, especially during the shoulder seasons (when Arctic snow cover exhibits the strongest variability) is of a special **interest.The** focus of this study is on the representation of the Arctic terrestrial snow in global circulation models from Coupled Model Intercomparison Project (CMIP5) ensemble during the melting (March-April) and the onset (October-November)	À cet égard, la représentation des rétroactions associées à la neige dans les modèles climatiques, en particulier pendant les saisons intermédiaires (lorsque la couverture neigeuse de l'Arctique présente la plus forte variabilité), est primordiale.Notre étude porte principalement sur la représentation de la neige terrestre arctique dans les modèles de circulation générale	Cette étude se concentre sur la représentation de la neige terrestre arctique dans les modèles de circulation générale de l'ensemble CMIP5 (Coupled Model Intercomparison Project) pendant la saison de fonte (mars-avril) et la saison d'apparition (octobre-novembre) pour la période allant de 1979 à 2005.

season for the period from 1979 to 2005.	issus du projet CMIP5 (Coupled Model Intercomparison Project) au cours du printemps (mars-avril) et de l'automne (octobre-novembre) de 1979 à 2005.	

Source EN	Reference FR	DeepL	OpenNMT
Snow characteristics from the general circulation models have been validated against in situ snow measurements, different satellite-based products and **reanalyses.We** found that snow characteristics in models have stronger bias in spring than in autumn.	Les caractéristiques de la neige des modèles de circulation générale ont été validées par rapport aux mesures de neige in situ, ainsi qu'à des produits satellitaires et à des réanalyses.Nous avons constaté que les caractéristiques de la neige dans les modèles ont un biais plus marqué au printemps qu'en automne.	Les caractéristiques de la neige provenant des modèles de circulation générale ont été validées par rapport aux mesures de neige in situ, à différents produits satellitaires et aux réanalyses.	Les caractéristiques de la neige issues des modèles de circulation générale ont été validées par rapport à des mesures in situ de la neige, de différents produits satellitaires et des réanalyses.

Source EN	Reference FR	DeepL
Even though the excavations were not extensive, many fossil specimens from a variety of large Pleistocene mammals were retrieved, including those of Reindeer, Horse, Red deer and **Rhinoceros.Best** documented in terms of both specimens and individuals is the steppe bison (Bison priscus).	Bien que le remplissage fossilifère n'ait été que faiblement entamé, de nombreux ossements furent exhumés se rapportant à plusieurs taxons de grands mammifères pléistocènes : Renne, Cheval, Cerf, Rhinocéros, ... mais le taxon le mieux représenté est le Bison des steppes tant en nombre de restes qu'en nombre d'individus.	Bien que les fouilles n'aient pas été très étendues, de nombreux spécimens fossiles d'une variété de grands mammifères du Pléistocène ont été récupérés, y compris ceux de rennes, de chevaux, de cerfs rouges et de rhinocéros.

Source EN	Reference FR	DeepL	OpenNMT
Climatic constraints on the alpine flora during the Younger Dryas oscillation and perhaps during other cold-climate events and intervening periods of higher temperature may have led to the loss of plant species in the White Mountain alpine **zone.Late-glacial** floras of lowland western New England were much richer than floras of areas above treeline during late-glacial time and at the present.	Les contraintes climatiques sur la flore alpine pendant l'oscillation du Dryas récent et peut-être aussi au cours d'autres périodes du tardiglaciaire ont peut-être contribué à la perte de certaines espèces de plantes dans la zone alpine des White Mountains.	Les contraintes climatiques exercées sur la flore alpine pendant l'oscillation du Younger Dryas et peut-être pendant d'autres épisodes de climat froid et des périodes intermédiaires de température plus élevée peuvent avoir entraîné la disparition d'espèces végétales dans la zone alpine des Montagnes Blanches.	Les contraintes climatiques sur la flore alpine lors de l'oscillation du Dryas récent et peut-être lors d'autres périodes de climat froid ou de température plus élevée ont peut-être conduit à la disparition d'espèces végétales dans la zone alpine du mont White.

Table 9 - Part of translation missing

Annex II: Automatic scores

Table 0 provides metric scores for all document types. Table 1 provides validation scores.

		Scores			
Type	**Engine**	**BLEU**	**TER**	**ChrF**	**COMET**
JAA	ModernMT baseline	33,18	60,23	62,54	81,84
	ModernMT OPERAS	33,52	59,96	62,72	81,99
	Deepl baseline	35,71	57,50	64,30	83,18
	Deepl termbase	35,50	57,82	64,08	83,07
	eTranslation	28,05	65,77	59,14	79,90
	OpenNMT baseline (30 epochs)	27,73	65,46	58,23	78,39
	OpenNMT OPERAS (30 epochs)	32,05	59,78	61,42	80,81
	OpenNMT OPERAS + SciPar (30 epochs)	32,55	59,96	60,88	80,65
	OpenNMT baseline (60 epochs)	28,76	64,00	58,79	78,98
	OpenNMT OPERAS (60 epochs)	31,14	61,26	60,47	80,25
	OpenNMT OPERAS + SciPar (60 epochs)	32,40	60,35	60,88	80,84
TA	ModernMT baseline	43,42	48,59	68,38	85,22
	ModernMT OPERAS	45,39	46,42	69,62	85,44
	Deepl baseline	45,60	46,82	69,62	86,06
	Deepl termbase	45,62	46,65	69,79	86,13
	eTranslation	35,16	55,47	64,25	83,63
	OpenNMT baseline (30 epochs)	37,40	52,85	65,14	82,89
	OpenNMT OPERAS (30 epochs)	40,33	49,14	66,98	84,18
	OpenNMT OPERAS + SciPar (30 epochs)	41,05	49,60	66,92	84,38
	OpenNMT baseline (60 epochs)	37,71	52,33	65,21	82,92
	OpenNMT OPERAS (60 epochs)	38,19	51,09	65,47	83,77
	OpenNMT OPERAS + SciPar (60 epochs)	41,91	48,29	67,18	84,49

Table 10 - Automatic scores for document types

Validation set	**OpenNMT**	**SacreBLEU**				
Engine	**10 epochs**	**20 epochs**	**30 epochs**	**40 epochs**	**50 epochs**	**60 epochs**
OpenNMT baseline	23.6	31.1	32.3	32.3	32.4	32.4
OpenNMT OPERAS	22.9	35.8	35.6	35.2	35.3	35.0
OpenNMT OPERAS + SciPar	28.7	36.8	36.8	37.3	37.5	37.7

Table 11 - BLEU score on validation set for every 10 iterations

Annex III: Adequacy task

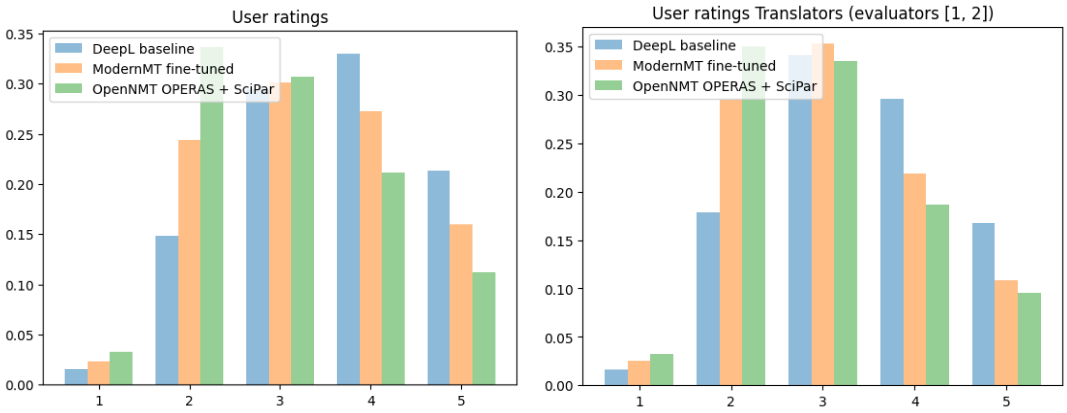
Setup and execution

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

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

Detailed results

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



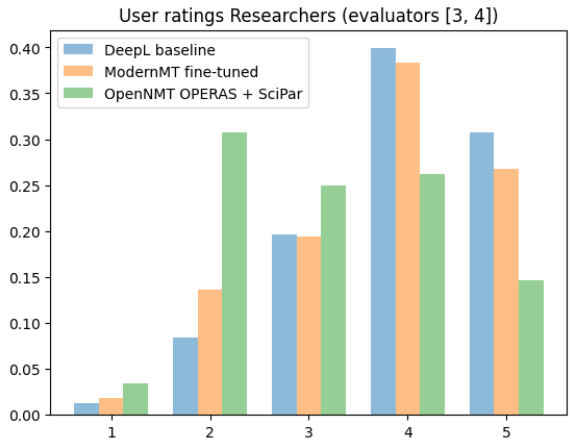


Figure 6 - User ratings per type of evaluator

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

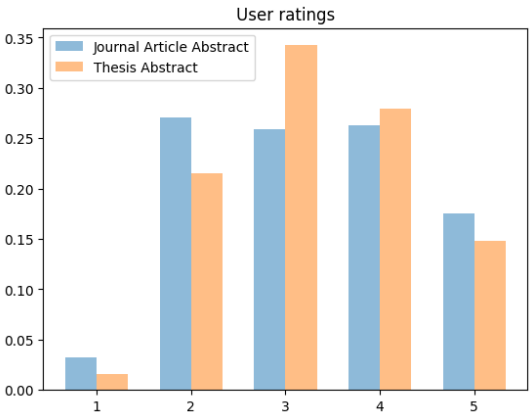


Figure 7 - User ratings per document type

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

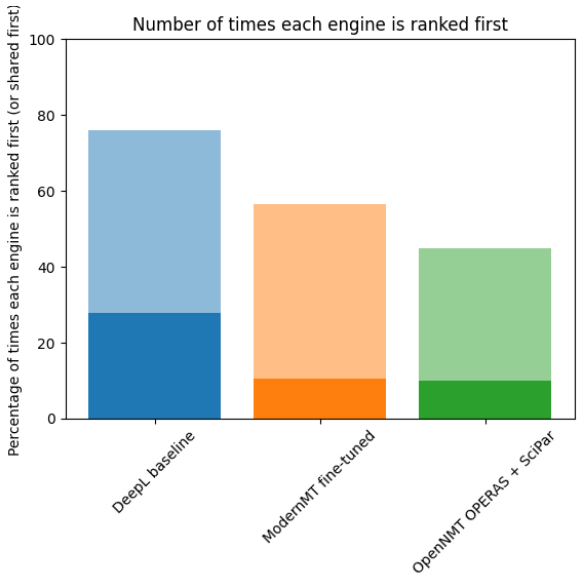


Figure 8 - Number of times engines are ranked first

ranked much more as sole best system than the other engines, and is also involved in many ties.

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

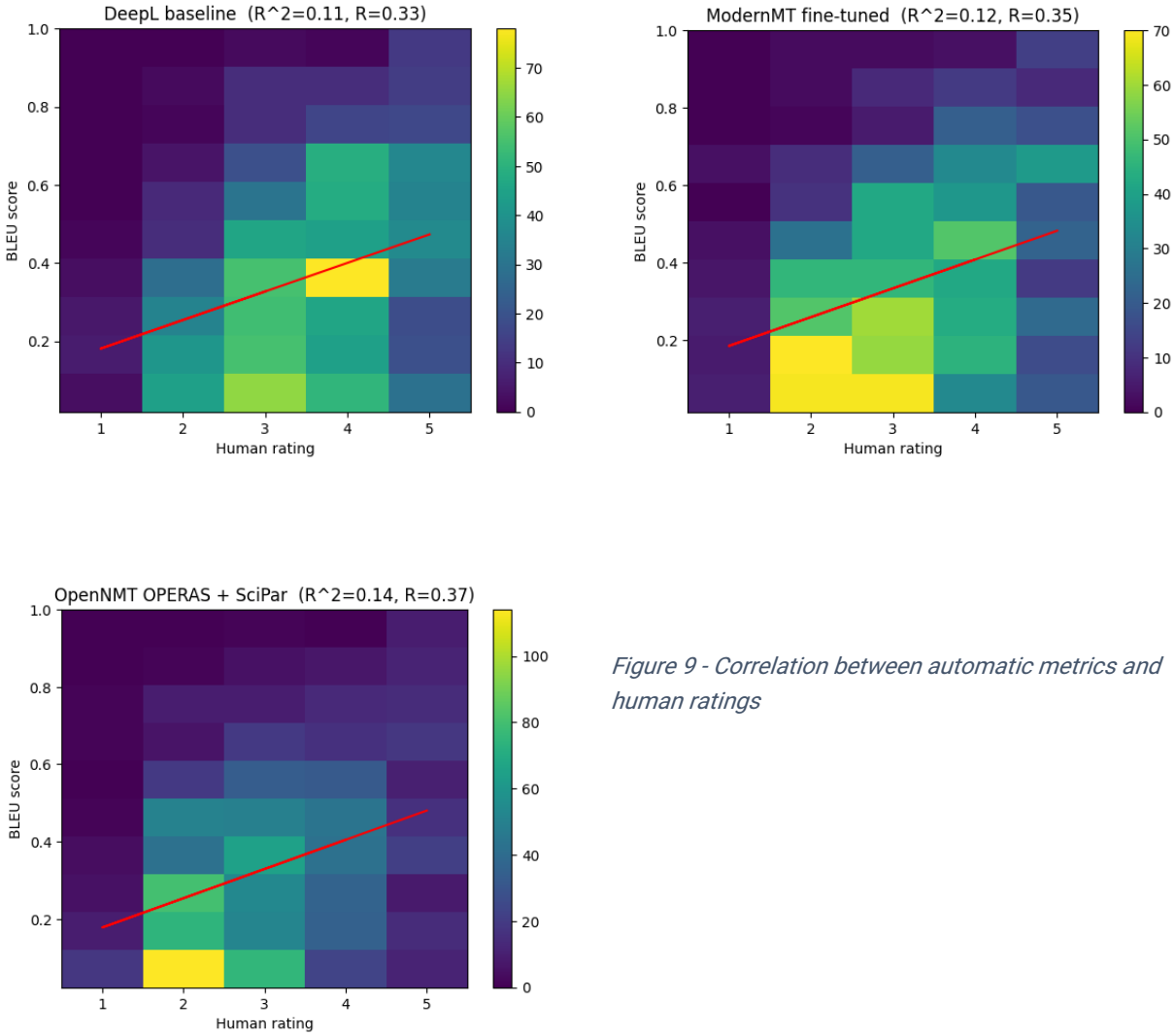


Figure 9 - Correlation between automatic metrics and human ratings

Annex IV: Productivity task

Setup and execution

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

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

Detailed results

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

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

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

SEM_Y = SEM(Y)

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

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

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

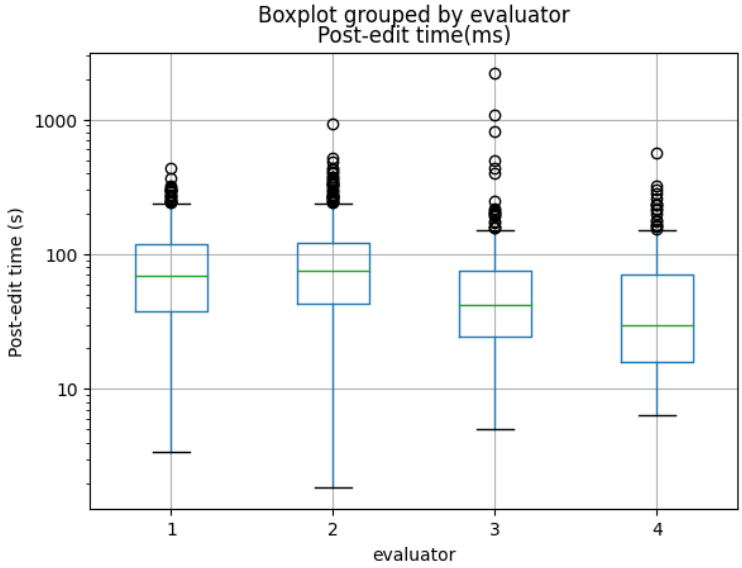


Figure 10 - Boxplot grouped by evaluator - post-edit time (ms)

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

Key takeaways:

- Even though each evaluator had a large difference in average post-edit time, the perceived effort still correlates well with post-edit time.
- We cannot say with significant confidence that the median post-edit times for a perceived effort of 4 and 5 differ.

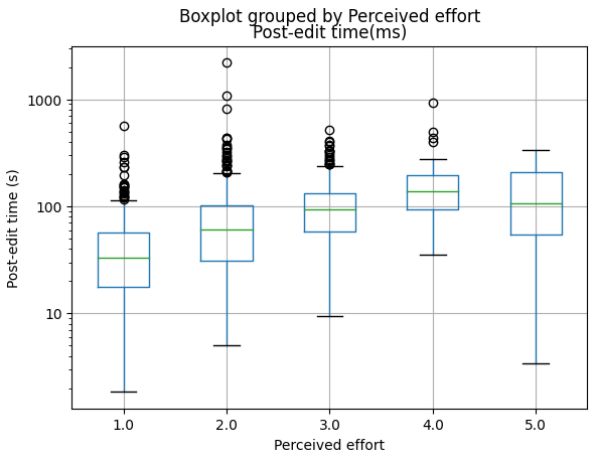


Figure 11 - Boxplot grouped by perceived effort - post-edit time (ms)

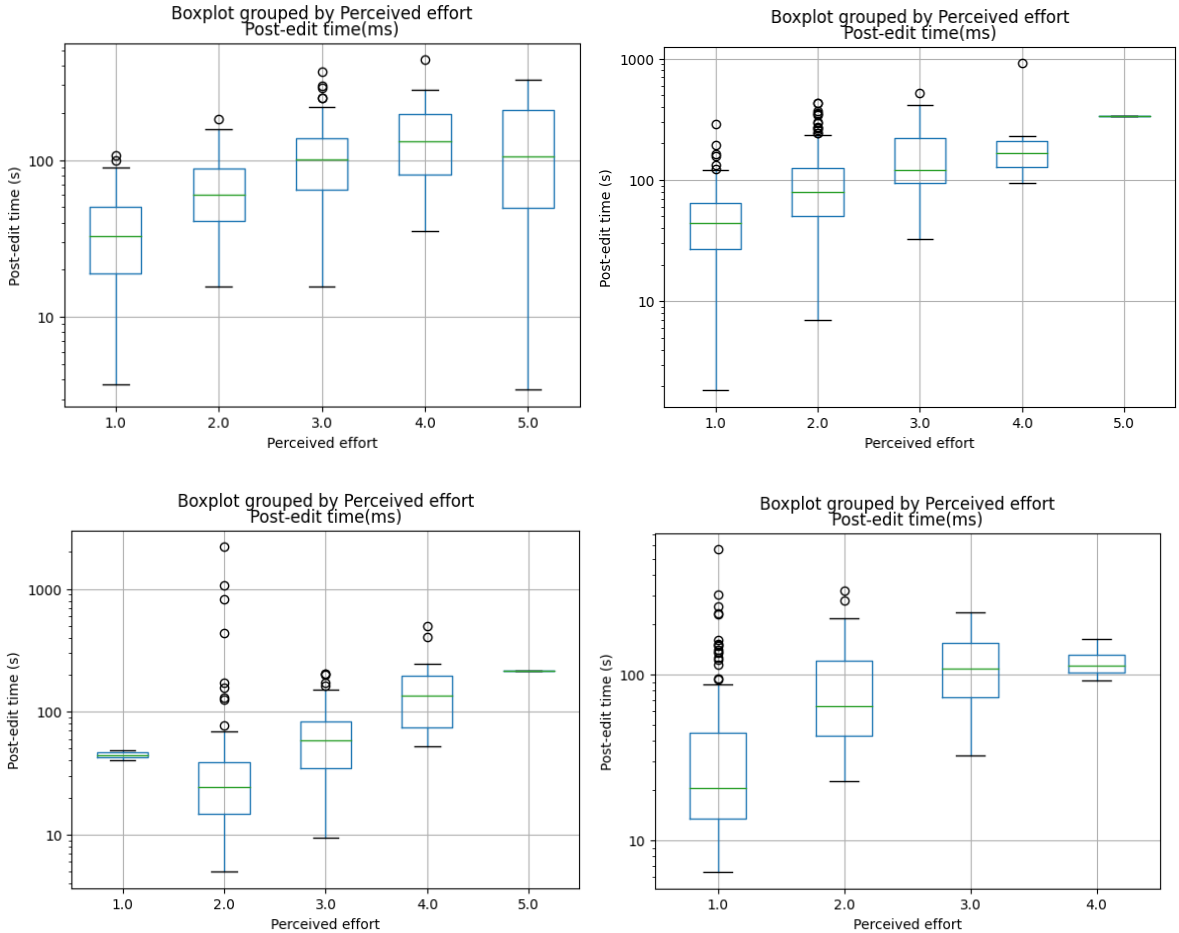


Figure 12 - Boxplot by perceived effort - post-edit time, individual evaluators

Figure 13 shows the post-edit time per engine. From the automatic evaluation we concluded that DeepL produces better outputs than ModernMT, and the latter, in turn, better outputs than OpenNMT. However, it appears that the post-edit times are almost identical between MT engines.

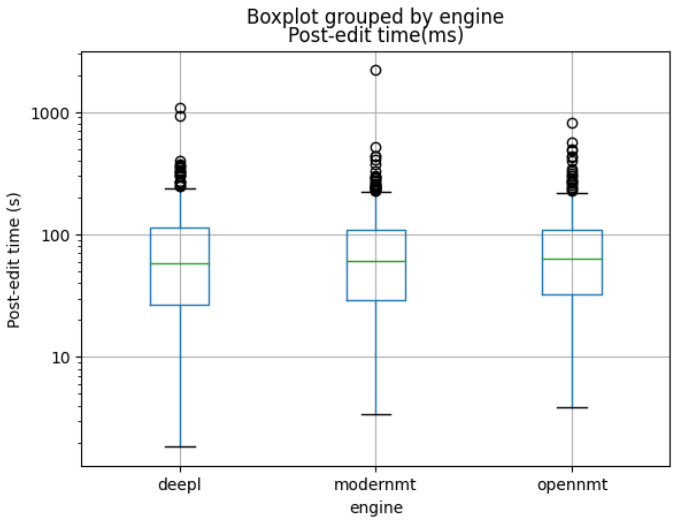


Figure 13 - Boxplot grouped by engine - post-edit time(ms)

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

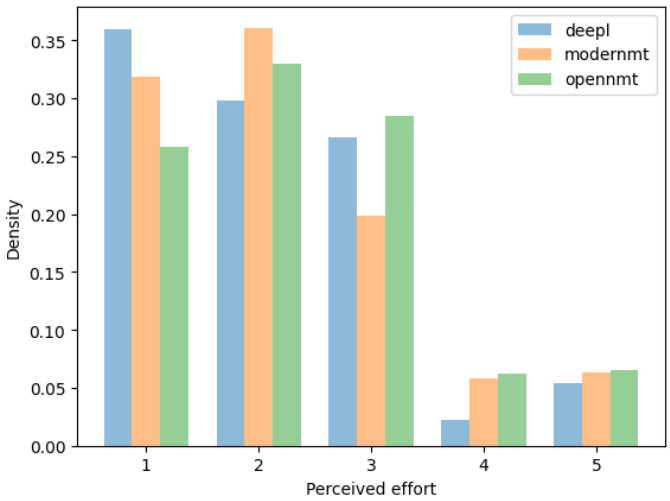


Figure 14 – Perceived effort per engine

Figure 15 shows the post-editing time per document type. There is hardly any difference between the journal article abstracts and thesis abstracts.

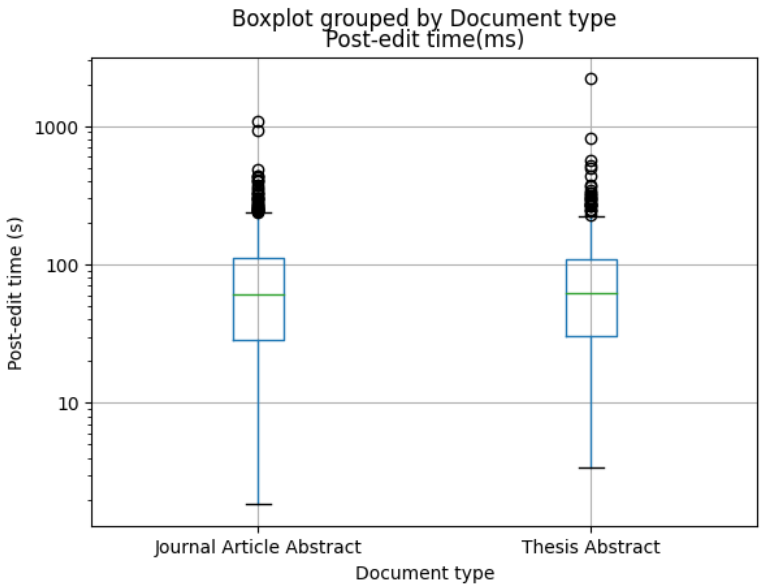


Figure 15 - Boxplot grouped by document type - post-edit time (ms)

The comparison of perceived efforts in Figure 16 confirms the previous findings. There is little difference between the two types of abstracts.

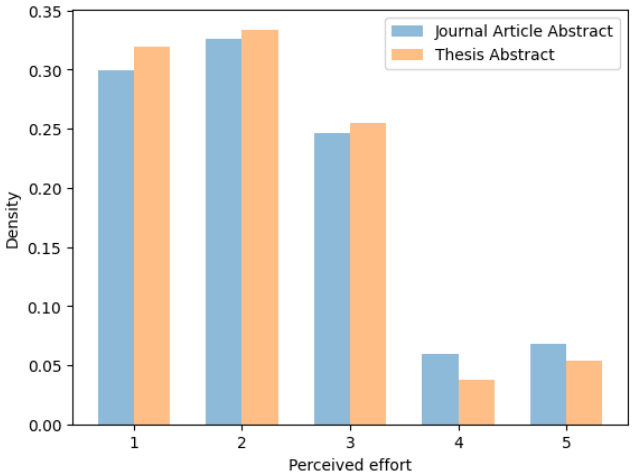


Figure 16 - Perceived effort per document type

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

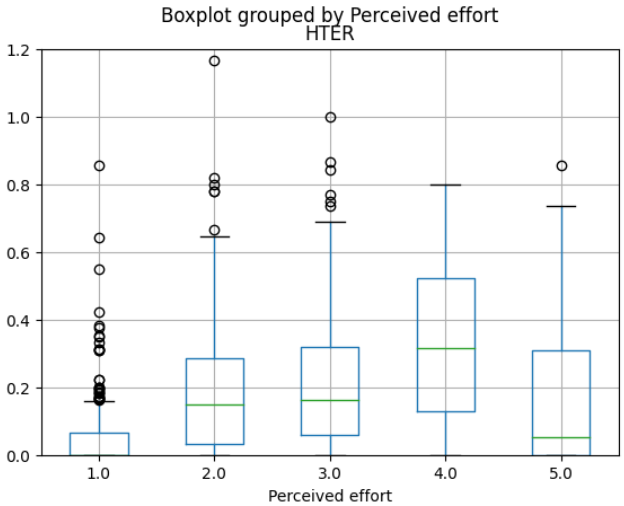


Figure 17 - Boxplot grouped by perceived effort - HTER

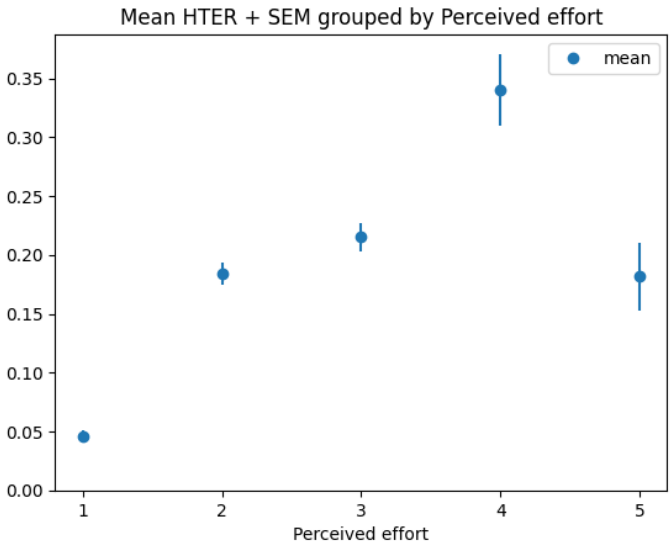


Figure 18 - Mean HTER + SEM grouped by perceived effort

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

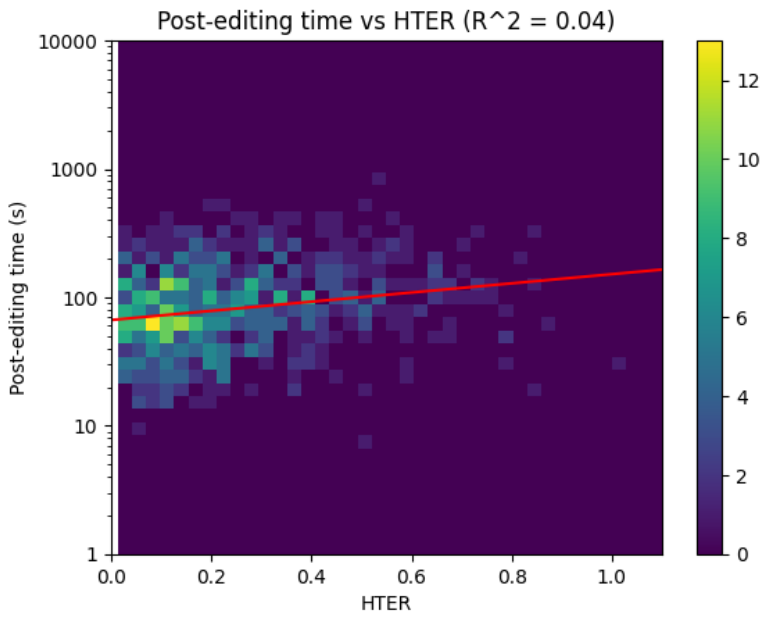


Figure 19 - Post-editing time vs HTER

Annex V: MQM error annotation results

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

Domain	MT System					
D3 ALL	**ModernMT**					
Error Severity Levels:	**Neutral**	**Minor**	**Major**	**Critical**	**Error Type Penalty Total**	
Severity Multipliers:	0	1	5	25		
Error Types	**Error Counts**				ET Weight	**ETPTs**
Term_Resource	9	3	3	0	1	18,0
Term_Inconsistent	1	0	0	1	1	25,0
Term_Wrong	0	0	1	4	1	105,0
Acc_Mistrans	0	4	**6**	**2**	1	84,0
Acc_Overtrans	0	0	0	0	1	0,0
Acc_Undertrans	0	1	0	0	1	1,0
Acc_Add	0	1	0	0	1	1,0
Acc_Omi	0	0	2	0	1	10,0
Acc_DNT	0	0	1	0	1	5,0
Acc_Untrans	0	0	0	0	1	0,0
Ling_Grammar	0	0	1	0	1	5,0
Ling_Punct	0	0	0	0	1	0,0
Ling_Spelling	0	1	0	0	1	1,0
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	1	5	3	0	1	20,0
Style_Unidimoatic	0	0	0	0	1	0,0
Style_Inconsistent	0	0	0	0	1	0,0
Loc_Number	1	0	0	0	1	0,0
Loc_Currency	0	0	0	0	1	0,0
Loc_Measure	0	0	0	0	1	0,0
Loc_Time	0	0	0	0	1	0,0
Loc_Date	0	0	0	0	1	0,0
Loc_Addr	0	0	0	0	1	0,0
Loc_Tel	0	0	0	0	1	0,0
Loc_Shortc	0	0	0	0	1	0,0
AudienceAppropriateness	0	0	0	0	1	0,0
DesignMarkup	0	0	0	0	1	0,0
			Absolute Penalty Total (APT):			275,00
Evaluation Word Count (EWC):	**1694**		**Per-Word Penalty Total (PWPT):**			0,1623
Reference Word Count (RWC):	1000		**Overall Normed Penalty Total (ONPT):**			162,34
Penalty Scaler (PS):	1,00		**Overall Quality Score (OQS):**			83,77
Max. Score Value (MSV):	100,00					
				Overall Quality Fraction		0,84
Total no. of errors	51			Sentences with errors		38,00
Total critical errors	7			Total sentences		62,00
				% Sentences with errors		0,61

Domain	MT System					
D3 ALL	**OpenNMT**					
Error Severity Levels:	**Neutral**	**Minor**	**Major**	**Critical**	**Error Type Penalty Total**	
Severity Multipliers:	0	1	5	25		
Error Types	**Error Counts**				ET Weight	**ETPTs**
Term_Resource	9	**5**	2	0	1	15,0
Term_Inconsistent	**0**	0	0	0	1	0,0
Term_Wrong	0	0	2	3	1	85,0
Acc_Mistrans	0	2	**10**	**3**	1	127,0
Acc_Overtrans	0	0	**2**	0	1	10,0
Acc_Undertrans	0	**0**	2	**0**	1	10,0
Acc_Add	0	0	0	0	1	0,0
Acc_Omi	0	1	4	1	1	46,0
Acc_DNT	0	0	1	2	1	55,0
Acc_Untrans	0	0	0	0	1	0,0
Ling_Grammar	0	0	5	0	1	25,0
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	1	5	2	0	1	15,0
Style_Unidimoatic	0	0	0	0	1	0,0
Style_Inconsistent	0	0	0	0	1	0,0
Loc_Number	1	0	6	0	1	30,0
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):			418,00
Evaluation Word Count (EWC):	1636		**Per-Word Penalty Total (PWPT):**			0,2555
Reference Word Count (RWC):	1000		**Overall Normed Penalty Total (ONPT):**			255,50
Penalty Scaler (PS):	1,00		**Overall Quality Score (OQS):**			74,45
Max. Score Value (MSV):	100,00					
				Overall Quality Fraction		0,74
Total no. of errors	69			Sentences with errors		46,00
Total critical errors	9			Total sentences		62,00
				% Sentences with errors		0,74

Domain	MT System					
D3 ALL	**DeepL**					
Error Severity Levels:	**Neutral**	**Minor**	**Major**	**Critical**	**Error Type Penalty Total**	
Severity Multipliers:	0	1	5	25		
Error Types	**Error Counts**				ET Weight	**ETPTs**
Term_Resource	11	4	1	0	1	9,0
Term_Inconsistent	0	0	0	0	1	0,0
Term_Wrong	0	0	2	4	1	110,0
Acc_Mistrans	0	2	**8**	0	1	42,0
Acc_Overtrans	0	0	**0**	0	1	0,0
Acc_Undertrans	0	1	0	0	1	1,0
Acc_Add	0	0	0	0	1	0,0
Acc_Omi	0	0	0	1	1	25,0
Acc_DNT	0	0	0	0	1	0,0
Acc_Untrans	0	0	0	0	1	0,0
Ling_Grammar	0	0	1	0	1	5,0
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	0	0	0	1	0,0
Style_Inconsistent	0	0	0	0	1	0,0
Loc_Number	0	0	0	0	1	0,0
Loc_Currency	0	0	0	0	1	0,0
Loc_Measure	0	0	0	0	1	0,0
Loc_Time	0	0	0	0	1	0,0
Loc_Date	0	0	0	0	1	0,0
Loc_Addr	0	0	0	0	1	0,0
Loc_Tel	0	0	0	0	1	0,0
Loc_Shortc	0	0	0	0	1	0,0
AudienceAppropriateness	0	0	0	0	1	0,0
DesignMarkup	0	0	0	0	1	0,0
			Absolute Penalty Total (APT):			204,00
Evaluation Word Count (EWC):	1658		**Per-Word Penalty Total (PWPT):**			0,1230
Reference Word Count (RWC):	1000		**Overall Normed Penalty Total (ONPT):**			123,04
Penalty Scaler (PS):	1,00		**Overall Quality Score (OQS):**			87,70
Max. Score Value (MSV):	100,00					
				Overall Quality Fraction		0,88
Total no. of errors	39			Sentences with errors		29,00
Total critical errors	5			Total sentences		62,00
				% Sentences with errors		0,47

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

CROSSLANG

www.crosslang.com

CrossLang NV
Amerikagebouw Kerkstraat
106 9050 Gentbrugge
Belgium

+ 32 9 335 22 00
info@crosslang.com