Translations and Open Science

Study on machine translation evaluation
in the context of scholarly communication

D1: Methodology for training, fine-tuning, collection, evaluation

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

Table of tables

Table of figures

1. Introduction

This document outlines the overall methodology followed for the study *Machine Translation evaluation in the context of scholarly communication*, carried out in the framework of the Translations and Open Science project (study corresponding to call 3 in the project).

In particular, the document describes the approach for training and fine-tuning (specialising) engines (open-source and commercial) for machine translation (MT), selecting appropriate testing material, automatically evaluating output of engines on the material (including baseline engines), manually evaluating output of engines while taking into account a number of personas, and annotating errors in the MT output using the MQM framework. Discipline-specific details and results are described in separate reports (D2, D3 and D4).

This document is structured as follows. In Section 2, we describe the approach for training, fine-tuning, and data selection. Section 3 describes the procedure for automatic evaluation. In Section 4, we explain the different types of human evaluation. Finally, we provide conclusions and references, as well a number of annexes relating to Sections 2 and 4.

2. Training and fine-tuning MT engines

2.1. Types of engines

We test both commercial engines and open-source software using MT engines which can be trained and fine-tuned. In case of commercial engines, we investigate the functionalities that allow for fine-tuning them (glossaries, translation memories). In case of open-source software, we train systems using various datasets, i.e. a baseline system based on generic data and systems based on both generic and discipline-specific data (note: an alternative setup, which may be more efficient in a production scenario, is to fine-tune a baseline model using discipline-specific data rather than training a discipline-specific model from scratch using a mix of generic and discipline-specific data).

We specifically perform tests with two commercial engines (DeepL and ModernMT) and one open-source software (OpenNMT), as described in Section 2.4. In addition, we also perform tests with eTranslation, the European Commission's MT tool (automatic evaluation only).[1]

2.2. Training and evaluation data

We make use of publicly available data for training a baseline model using open-source software. For fine-tuning commercial engines and training discipline-specific models using open-source software, we use scientific publications and abstracts available on web pages, as well as bilingual terminology lists.

We specifically make use of the data extracted from web pages in the study *Mapping and collection of scientific bilingual corpora* (study corresponding to call 1 in the project). The data consists of several publication types (such as Journal Articles, Journal Article Abstracts and Thesis Abstracts) and their metadata (document ID, sizes, domain, publication type, publication source, URL, title, keywords, author and license), as well as a terminology list. For each discipline, around 100k sentence pairs and a term base of 300 terms are available. As described in Section 2.3, we use most of the data provided as training set, and, from the remaining data, extract a validation set (to be used during training), a test set (for automatic evaluation and selection of the final model) and an evaluation set (for human evaluation). In addition, with the support of OPERAS, we look for additional data for human evaluation, in the first place to have additional scientific sources of information compared to the training set, and in the second place to have sufficient popularising material, in case the material at hand has an exceedingly high scientific level for a broad audience.

The data from call 1 was provided in the form of segment list, including metadata about the document source. These needed to be processed. They were grouped per document and parallel source and target documents were created. The metadata was saved separately at document level. The script for processing the segment list is provided here: https://github.com/CrossLangNV/Translations-and-Open-Science.

[1] https://commission.europa.eu/resources-partners/etranslation_en

2.3. Data partitioning

The collected data is partitioned into the following types of subsets:

1. **Training set**
   The set of examples used to train a custom MT model. Based on this data, the model will learn to make predictions on new unseen data.

2. **Validation set**
   A small subset held out from the training data used to monitor the performance of the model during training of the open-source MT system and allowing to determine when to stop training to prevent "overfitting" of the model.[2] The validation subset does not overlap with the test set in order to separate the development process from the actual evaluation of the system's performance. It is important not to look at the test and (human) evaluation sets while still developing or improving the model.

3. **Test set**
   A set of unseen examples which is used to evaluate the performance of an MT model (in case of the open-source MT system, the model determined by the validation set) and select the final MT models for (human) evaluation. This set should therefore be separated from the datasets used for (human) evaluation. Test sets are scored using automated metrics only. Potential problems with MT models are typically detected by looking at the translation output of a model and looking for mistakes that seem to be recurring and follow a certain pattern.

4. **Evaluation set**
   The actual evaluation data used for human evaluation are a mix of "internal" (i.e. held out from the original dataset) and "external" (i.e. coming from new sources) texts in order to assess how well the selected models perform on similar and new data. By including both internal and external data sources, the evaluation process addresses the generalizability and robustness of the MT models (data sparsity). Therefore, and depending on the objectives and requirements of the human evaluation, the evaluation set may have a different distribution than the training, validation and test data. Additional selection criteria may apply, such as text complexity (e.g. for the self-paced reading experiments) and license type (open licenses are preferred). In order to measure the correlation between human and automatic evaluation results, the evaluation sets will also be scored using automatic metrics. Therefore, it is

[2] Overfitting occurs when an MT model is trained to fit the new domain too closely so that it becomes unable to generalize well to new unseen data.

preferred to use parallel evaluation sets that include both the source and reference translations, even if the reference translations are not directly used during the human evaluation experiments. In order to facilitate the establishment of human evaluation tasks, we set up an evaluation set that is sufficiently large to sample from, as each task has its own focus; the selected samples may vary across tasks (see the above criteria) but also partially overlap, in order to allow for calculating cross-task statistics (e.g. correlation between human evaluation task scores).

In addition, we also adhere to the following general principles during the data partitioning process:

- Validation and test sets should have a similar distribution to the training data, particularly with regards to the text types (such as journal articles, journal article abstracts, and thesis abstracts), and should include different sources, if possible. Simultaneously, each type should have a large enough sample size to be statistically significant.
- Documents should not be divided between different subsets. In other words, each text should be kept intact and not split between training, validation, test and evaluation datasets.
- An important consideration for the composition of the evaluation set was the open license, allowing the data to be shared with translators for human evaluation purposes, as well as the potential open sourcing of the data at a later time.
- The trained open-source systems are tested for regression: to ensure that the system is not deteriorating in comparison to the baseline model (i.e. overfitting on the in-domain data), additional "generic domain" test sets are used and scored using automatic metrics.

The script for ensuring the distribution of the validation and test sets is provided on the GitHub site mentioned earlier.

Detailed dataset statistics are provided in the discipline-specific reports (D2 to D4).

2.4. MT Customisations

For each discipline, we perform a set of MT customisation experiments using two commercial systems that are widespread among the translation community (DeepL and ModernMT) and one open-source solution (OpenNMT). The translation direction of the models depends on the discipline: for *Neuroscience and Disorders of the Nervous System* and *Climatology and climate change*, this is English to French, *for Human Mobility, Environment, and Space*, this is French to English.

- **DeepL translator:** DeepL[3] is a commercial MT provider offering users the ability to customise translations by adding a glossary (list of terms)[4] and enforcing a translation of a particular term, potentially resulting in better consistency of translated terms and, consequently, better quality translations. DeepL does not offer the option to customise a model using a parallel training dataset. We use a DeepL Pro Classic subscription for €16.53/month. The subscription includes 1 million free characters per month. Translations beyond that are charged €20 per 1 million characters.

- **ModernMT:** The second commercial system, ModernMT,[5] offers the option to customise a generic model by simply uploading a user translation memory (TMX) without the need for a traditional pre-training process (the creators refer to this as "instance-based adaptation"). We make use of a ModernMT Batch license. This is a pay-per-use license with a rate of $8 per 1 million characters to be translated. It should be noted that our license does not include the option to upload documents and comes with a maximum limit on the length of the test text that can be translated at once, requiring several copy-paste actions given a test set of a substantial size.

- **OpenNMT:** OpenNMT[6] is an open-source toolkit allowing developers to build and customise their own MT systems using pre-processing, training, and translation tools. For each discipline, we train three models. As a baseline, we train an MT model from scratch on publicly available data from the OPUS repository (Tiedemann 2009).[7] In parallel, we train an in-domain MT model adding the domain-specific training data on top. A third model uses the SciPar[8] dataset in addition, to see whether and to what extent similar (pseudo in-domain) data might help further improve the quality of the translations. For all three models, we use the same hyperparameters for training. We apply BPE segmentation (Sennrich et al. 2016) with 32k merge operations (about 32k subtokens joint vocabulary). We train on sample sizes of 2.5M for 50 epochs, using the transformer_big training parameters. To make sure the OpenNMT models do not tend to overfit on the targeted domain, we take the following measures:
    - During training, a validation set is used allowing us to monitor the model's potential overfitting on in-domain data and to decide when to stop training;

[3] https://www.deepl.com/translator

[4] Note that this option is only available for a limited number of languages, i.e. English, German, French, Dutch, Spanish, Italian, Japanese and Polish. A maximum of 5000 entries is allowed per glossary.

[5] www.modernmt.com

[6] We use the OpenNMT TensorFlow implementation (https://github.com/OpenNMT/OpenNMT-tf)

[7] https://opus.nlpl.eu. See Annex I for an overview of the datasets used.

[8] The SciPar corpus is available via OPUS and consists of about 1M abstracts (https://aclanthology.org/2022.lrec-1.284/).

- During automatic evaluation, scores for generic domain test sets are monitored as a significant decrease in quality might indicate the model's potential overfitting.

3. Automated evaluation

We evaluate the quality of the engines trained using the procedure in Section 2.4 by applying the engines to the source segments in the test set described in Section 2.3, as well as to subsets of it (based on document type), and running automated metrics to the output and the reference translation in the test set. We also visualise samples of differences between MT output. By applying these metrics and visualisations, a decision can be taken as to which engines will be submitted for human evaluation (see Section 4): if a specialised engine performs much lower in various aspects than others, it may be decided to submit merely the latter engines to the human evaluation tasks.

To automatically measure the quality of the translation outputs compared to their reference translation, we compute different standard MT metrics for each system, such as BLEU (Bilingual Evaluation Understudy, Papineni et al. 2002) and TER (Translation Edit Rate),[9] as well as neural metrics such as COMET (Rei et al., 2020). We also created reports for each type of system (produced using CrossLang's MT Advisory tool) where we compare baseline versus customised models, showing BLEU score and edit distance at segment level, as well as a comparison view with the edits needed to obtain the reference translation; see the example in Figure 1.

Type	Sentence	dist	BLEU
SRC	Ainsi, pour la plupart des habitants, quelqu'un d'extérieur à la région qui contribue à préserver et enrichir le territoire ne sera pas forcément rejeté.	-	-
REF	For most of the inhabitants, therefore, someone from outside the region who helps to preserve and enrich the territory will not necessarily be rejected.	-	-
modernmt baseline	Thus, for most inhabitants, someone from outside the region who helps preserve and enrich the territory will not necessarily be rejected.	28	0.6546
REF/modernmt baseline	~~Thus, I~~For most of the inhabitants, therefore, someone from outside the region who helps to preserve and enrich the territory will not necessarily be rejected.	28	0.6546
modernmt	Thus, for most inhabitants, someone from outside the region who helps to preserve and enrich the territory will not necessarily be rejected.	25	0.7617
REF/modernmt	~~Thus, I~~For most of the inhabitants, therefore, someone from outside the region who helps to preserve and enrich the territory will not necessarily be rejected.	25	0.7617

Figure 1 - Excerpt from report showing sentence-level scores and a comparison view of the different outputs with regards to the reference translation

For the automatic evaluation, we make use of the test sets outlined in Section 2.3. Additionally, we create parametrized subsets by dividing the test sets based on their text type, allowing us to identify potential differences in performance between various types of documents, such as journal articles, journal article abstracts and thesis abstracts.

[9] We use the SacreBLEU implementation.

In order to get a better idea of the score differences between MT engines, we calculate the statistical significance of each metric. Calculating the standard error of the mean (SEM) gives us an idea of the certainty interval in which the true mean score is contained.[10]

[10] See https://en.wikipedia.org/wiki/Standard_error#:~:text=Standard%20error%20of%20the%20sample%20mean%5Bedit%5D

4. Human evaluation

The specialised engines considered to be sufficiently valuable according to the procedure described in Section 3 are submitted for human evaluation. The latter is undertaken from three different perspectives (profiles), i.e. translator, expert and layperson, and consists of three tasks, i.e. the adequacy task, the post-editing task, and self-paced reading experiments. In these tasks, subsets of the evaluation set (see Section 2.3) and the MT output from the different engines are used in a variety of ways. As part of the third task, we also undertake error annotations to get a better view of the relation between human judgment and MT errors.

4.1. Profiles

We consider the following profiles:

1. "translator": professional translator who masters the source language, is a native speaker of the target language, and has a good knowledge of the discipline in question. This persona performs post-editing in an environment such as a computer-aided translation (CAT) tool.
2. "expert": researcher specialised in a discipline and applying MT to (a) translate his/her scientific publication, (b) write an article in the target language (writing aid), or (c) perform gisting of scientific texts that are not written in the native language (reading aid). In the first two cases, the knowledge of the target language should be sufficient to be able to judge adequacy and grammaticality of discipline-specific text in that language.
3. "layperson": a person who has at most basic knowledge in the discipline (e.g. a non-academic person or a researcher in a different scientific discipline). This persona has good/excellent knowledge of the target language and makes use of MT to perform gisting of popularising scientific texts.

4.2. Setup of paragraph samples

In order to prepare the tasks for the evaluators, we take samples from the evaluation set described in Section 2.3. This set is sufficiently large to enable the selection of source segment samples appropriate for the tasks at hand. The samples for three tasks satisfy the following constraints:

- They consist of paragraphs taken from articles or abstracts (actual or artificial paragraphs, as described below).
- The samples for the adequacy task and the post-editing task are disjointed, in order to avoid bias (one evaluator may perform both tasks and should only evaluate one specific segment once).
- The sample for the adequacy task and post-editing task covers various document types and contains segments which may be highly technical as well as popularising. The sample for the

self-paced reading experiments focuses on popularising text or text which may be technical but still sufficiently understandable for a larger audience.

- There may be overlap between the sample of the adequacy or post-editing task and the one of the self-paced reading experiments. This allows for comparing the scores among various tasks.
- Samples should have a similar distribution in terms of automatic evaluation scores compared to the full evaluation set.

In order to determine paragraphs in the evaluation set, we manually indicate paragraph boundaries in it (the set is loaded into Excel and the start of a paragraph is flagged on the row of the segment in question). This manual process is not applied to abstracts in the evaluation set, as they typically consist of one block of text. Rather, this block of text is split automatically into artificial paragraphs.

4.3. Adequacy task

In this task, the evaluator (translator or researcher) judges the adequacy of translated segments (of type sentence) in the MT output of a specialised text, by assigning a score (rating) between 1 and 5; several MT outputs are shown per source segment, so the assignment of scores implicitly provides a ranking. The aim of this task is to assess how adequately the translation of the segment expresses the source segment's meaning, and, by consequence, how useful the translation is for gisting.

The adequacy task is set up as follows:

- Randomly select paragraphs from the evaluation set.
- List the selected paragraphs in a random order.
- Translate the segments in the selected paragraphs using each of the three MT engines.
- Manually check the segments and their MT output, as well as their reference translations (see below for types of comparison of automatic scores and human judgment). The source segment should correspond to the reference translation, and the MT outputs should correspond to the source segment; as data collection and processing involve various automated steps, unexpected results from these steps cannot be excluded. In case of an unexpected result, assess the potential solution on a case-by-case basis (e.g. resubmit source segment to MT engine).
- Convert the resulting source segments and their MT outputs to input files for CrossLang's Machine Translation Evaluation (MT-Eval) tool (CSV format). Each input file constitutes a "batch" in MT-Eval; it contains around 100 segments. Batches are used in order to keep the evaluation process manageable (e.g. in case something goes wrong with a file in MT-Eval, only a small part of the work needs to be redone).

MT-Eval shows the following information for each segment:

1. source segment

2. MT outputs (randomly ordered by MT-Eval, to avoid evaluator bias), each one followed by a set of labelled numbers to choose from.

Figure 2 shows a screenshot of a segment to be evaluated using MT-Eval.

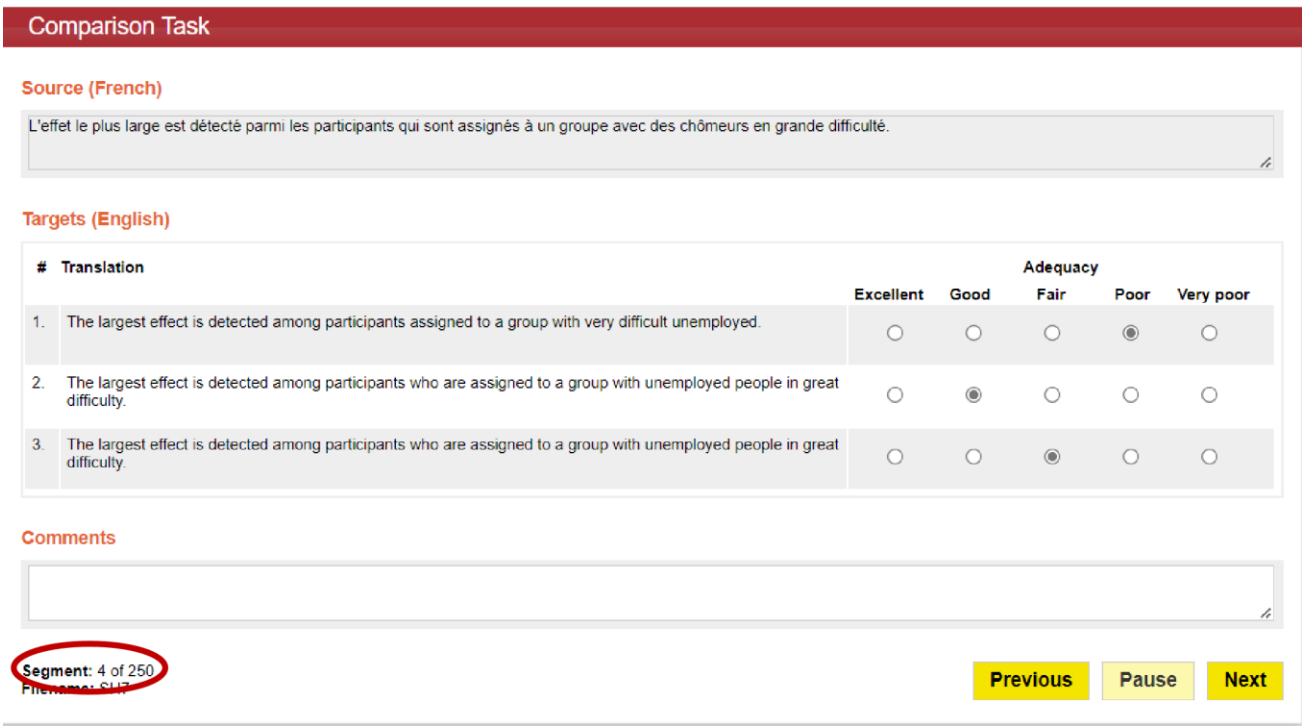


Figure 2 - Comparison Task – CrossLang Machine Translation Evaluation tool

Reference translations are not shown (again, to avoid bias).

For each discipline, the task is performed by four to six evaluators (including translators and researchers), who each evaluate the MT output for around 500 source text segments. The contact with potential evaluators was established with the support of the University of Rennes and of OPERAS. The financial compensation proposed to potential evaluators either consisted of (a) a price based on an estimated 1-minute manual processing time per source segment and a fixed price per hour or (b) a fixed price for the whole task. The proposed time span to finalise the task (which also applied in case the evaluator performed two tasks, i.e. also the post-editing task discussed in Section 4.4) consisted of two weeks. Upon agreement of the proposed compensation and time frame by the evaluator, the following information was sent to the evaluator:

1. Instructions for performing the adequacy task, with a matrix explaining how to assess the conveyance of meaning from source segment to MT output (see Annex III).
2. Links for opening the batches in MT-Eval.
3. CrossLang's standard NDA, to be signed by the evaluator.

4. (In case the evaluator is a researcher) a simple service contract, including information on the price and deadline, to be signed by the researcher.
5. Background information in the form of (a) the terminology list provided by call 1 of the Translations and Open Science project and (b) the abstracts from which the segments to be evaluated originate from.

Based on the evaluation outcome (enriched CSV files), a number of statistics can be produced:

- Distribution of ratings for MT engines
- Distribution of ratings for document types
- Number of times an engine is ranked first
- Correlations between automatic scores (e.g. BLEU) and human ratings

The above statistics can also be calculated separately per type of evaluator (in order to compare ratings / rankings between translators and researchers).

4.4. Post-editing task

In this task, the evaluator (translator or researcher) is asked to post-edit translated segments in order to obtain a publishable translation.

The post-editing task is set up in the same way as the adequacy task, with the following differences:

- We add the string “NEW PARAGRAPH” to the first segment of a paragraph to indicate to the evaluator which segments belong together.
- We translate each segment using just one MT engine, which is arbitrarily determined for a segment, such that MT engines are more or less equally distributed across the MT outputs proposed to the evaluator.
- When creating batch files for MT-Eval, we ensure each segment is preceded by the previous segment in the paragraph and followed by the next one (if the segment is the start or end of the paragraph, the preceding/following segment belongs to another paragraph) and that each segment is followed by a dummy segment which allows for specifying perceived post-editing effort. This context provides support to the evaluator when translating a segment.

The batch files allow MT-Eval to show the following information for each source segment in it:

1. preceding source segment
2. its MT output
3. source segment to translate
4. its MT output
5. next source segment

Figure 3 shows a screenshot of a segment to be evaluated using MT-Eval:

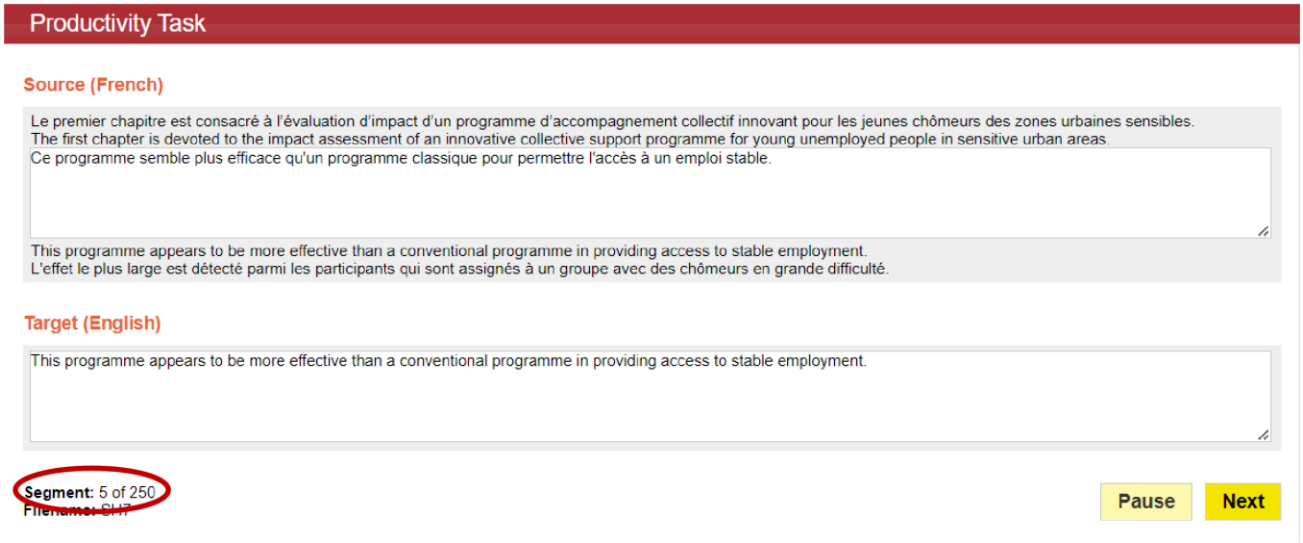


Figure 3 - Productivity Task – example segment to be evaluated in MT-Eval

Figure 4 shows a screenshot of a dummy segment:

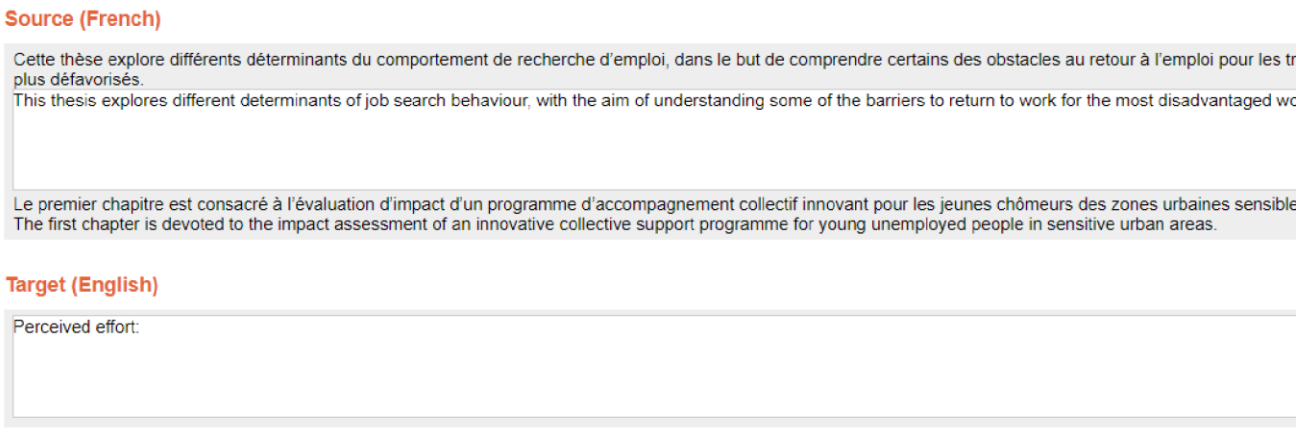


Figure 4 - Productivity Task – example of a dummy segment

For each discipline, the task is performed by four to six evaluators (including translators and researchers), who each post-edit the MT output for around 500 source text segments. The same procedure for contacting evaluators and price and time span proposal was followed as in case of the adequacy task; however, the first price option (i.e. the one not consisting of a fixed price) consisted of estimating the number of hours based on the average sentence length of the segments involved and a post-editing speed of 750 words per hour (after consultation with University of Rennes). The following information was sent to the evaluator:

1. Instructions for performing the post-editing task, with an explanation of the criteria to be fulfilled by a publishable translation (see Annex IV).
2. Links for opening the batches in MT-Eval.

3. (Unless the evaluator also performed an adequacy task and already signed the document) CrossLang's standard NDA, to be signed by the evaluator.
4. (In case the evaluator is a researcher and did not perform an adequacy task yet) a simple service contract, including information on the price and deadline, to be signed by the researcher.

As in case of the adequacy task, background information in the form of a terminology list and abstracts is provided.

Based on the evaluation outcome (enriched CSV files), a number of statistics can be produced:

- Distribution of post-edition time (per segment) across evaluators
- Correlation between post-edition time and perceived effort
- Correlation between post-edition time and MT engine
- Correlation between perceived effort and MT engine
- Distribution of post-edition time across document types
- Relation between HTER (human-target TER) and perceived effort
- Correlation between HTER and post-edition time

4.5. Self-paced reading experiment

In this experiment, the evaluator is a layperson. We selected short text excerpts of 120-200 words from the evaluation set. Based on text characteristics, e.g. the origin of the excerpt (abstract or full text), sentence length and text difficulty in combination with the automatic evaluation scores, the texts were selected and classified into different sets.

For each set, we made sure that all conditions (human reference translation, the output of the first MT system, the output of the second MT system and the output of the third MT system), were evenly distributed (see the balanced design provided in Table 1).

At least three participants read all texts of one set. The participants were highly proficient in the target language.

The total duration of the experiment was limited to one hour. The experiment started with a practice text to familiarise the participants with the task. Participants read the texts (in an isolated room, with minimal distraction) in a cumulative self-paced reading view, in which each button press reveals the next sentence while the previous sentences of the text stay in view. After each excerpt, a yes/no comprehension question was asked to give the participants an incentive to read the text attentively. After each excerpt, the participants answered the question whether the translation quality was sufficient to get an idea of the scientific text.

	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 1 - Example of the experimental design

The self-paced reading experiments were executed using the Zep Experiment Control Application (version 2),[11] a toolkit used for experimental psycholinguistics. Reading time was measured per sentence.

4.6. MQM Error Annotation

All texts used in the self-paced reading experiments are annotated for errors using the MQM framework.[12] Prior to annotating MT errors, terms are marked in the source texts: (i) using the corresponding terminology list provided per discipline, and (ii) by the annotator. When using the term lists, terms are marked automatically (custom Python code). This automatic process works on lemmatized and lowercase text (except when entries are fully uppercase). In case of manual marking, the term extraction methodology proposed by Rigouts Terryn et al. (2020) is applied.

Error annotation is performed by a single annotator, using the MQM error types, details of which are provided in Annex II: terminology, accuracy, linguistic conventions, style, locale conventions, audience appropriateness, and design and markup. For each MQM error type, fine-grained error annotations are made using the corresponding sub-categories.[13]

The MQM decision tree[14] is used to decide on the type of translation error, which is detected on a specific text span.

[11] https://www.beexy.nl/zep/wiki/doku.php?id=home

[12] https://themqm.org/

[13] https://themqm.info/typology/

[14] https://themqm.org/error-types-2/decisiontree/

MQM Scorecard templates[15] are used to assign annotated errors to numerical values based on severity levels and corresponding severity multipliers. These values provide the data needed to "measure" translation quality in terms of the MQM metric. Four severity levels are considered: neutral, minor, major, and critical. Using the scorecard templates, four calculations (Error Count, Absolute Penalty Total, Per Word Penalty Total, Overall Quality Score) are made to convert error annotations to quality scores on sentence and document levels. Details on the severity levels and the calculations are provided in Annex II. In addition to the MQM scores, the ratio of the number of sentences with MT errors, total number of critical errors and total number of errors are calculated per engine.

The error annotation task is performed using the Label Studio Toolkit[16] (the interface of which is illustrated in Figure 5), with a local installation. The annotation task is performed per source sentence with the outputs listed for all three engines line by line. The order of the MT engines is randomised between different files (but are kept the same within the same file to potentially capture discourse-related errors throughout a given text). The analysis of the error annotations is automatically performed using the JSON export files obtained from Label Studio.

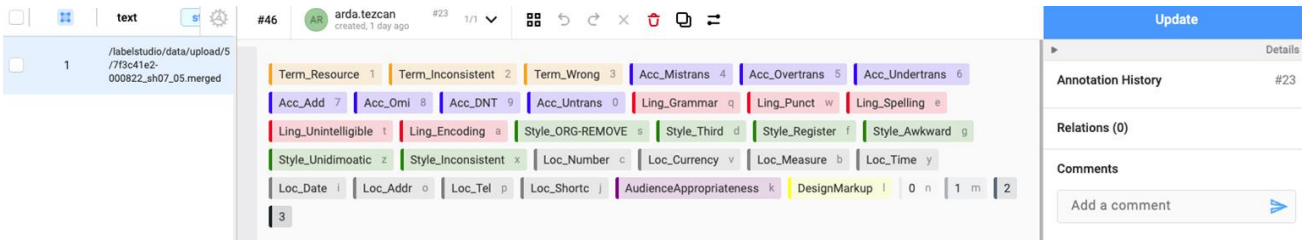


Figure 5 - Interface of the Label Studio Toolkit

[15] https://themqm.org/error-types-2/1_scorecards/design/

[16] https://labelstud.io/

5. Conclusions

In this document, we outlined the overall methodology for our MT evaluation study. We described the types of MT engines, the approach for training and fine-tuning them, the procedure for data selection, the method for automatic evaluation, and the different types of human evaluation we applied.

Reports D2 to D4 provide details on the application of the methodology to three predetermined scientific disciplines. The present document will allow the methodology to be replicated in the future on additional scientific disciplines and in an environment that is subject to production constraints.

6. References

Papineni, Kishore, Salim Roukos, Todd Ward, and WeiJing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pages 311–318.

Rei, Ricardo, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A Neural Framework for MT Evaluation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 2685–2702.

Rigouts Terryn, Ayla, Véronique Hoste, and Els Lefever. "In no uncertain terms: a dataset for monolingual and multilingual automatic term extraction from comparable corpora." Language Resources and Evaluation 54, no. 2 (2020): 385–418.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016), pages 1715–1725.

Tiedemann Jörg. 2009. News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In Recent Advances in Natural Language Processing (Vol. V), pages 237–248, John Benjamins, Amsterdam/Philadelphia.

Annex I: Out-of-domain datasets

Domain	Sample distribution	Corpus name	Corpus size
Dialog	0.17	EUbookshop	10617855
		OpenSubtitles	41762988
Medical	0.17	EMEA	1092068
Legal	0.17	DGT	4938565
		ELITR-ECA	441081
		JRC-Acquis	813667
News	0.17	News-Commentary	155622
Education	0.16	QED	789911
Patent	0.16	EuroPat	11098211

Table 2 - Out-of-domain datasets

Annex II: Details on MQM framework

High-level error dimensions:

- **Terminology** – errors arising when a term does not conform to a normative domain or organizational terminology standards or when a term in the target text is not the correct, normative equivalent of the corresponding term in the source text.
- **Accuracy** – errors occurring when the target text does not accurately correspond to the propositional content of the source text, introduced by distorting, omitting, or adding to the message.
- **Linguistic conventions (Fluency in version 1)** – errors related to the linguistic well-formedness of the text, including problems with grammaticality, spelling, punctuation, and mechanical correctness.
- **Style** – errors occurring in a text that are grammatically acceptable but are inappropriate because they deviate from organizational style guides or exhibit inappropriate language style.
- **Locale conventions** – errors occurring when the translation product violates locale-specific content or formatting requirements for data elements.
- **Audience appropriateness (Verity in version 1)** – errors arising from the use of content in the translation product that is invalid or inappropriate for the target locale or target audience.
- **Design and markup** – errors related to the physical design or presentation of a translation product, including character, paragraph, and UI element formatting and markup, integration of text with graphical elements, and overall page or window layout.

Severity levels:

1. **Neutral: severity multiplier 0 (no effect on final error score)**

   In this case, the evaluator considers that a different solution is warranted, but that the translation should not be penalized for an error. For instance, the root cause may be beyond the translator's control, a termbase may have been incorrect, the evaluator's suggested change is only preferential, or the severity of the error does not warrant even minor severity. This value can be used to flag items for fine-tuning feedback purposes.

2. **Minor - severity multiplier 1**

A minor error instance has a limited impact on, for example, accuracy, stylistic quality, consistency, fluency, clarity, or general appeal of the content, but it does not seriously impede the usability, understandability, or reliability of the content for its intended purpose.

3. **Major - severity multiplier 5**

   A major error instance seriously affects the understandability, reliability, or usability of the content for its intended purpose or hinders the proper use of the product or service, for instance due to a significant loss or change in meaning or because the error appears in a highly visible or important part of the content.

4. **Critical - severity multiplier 25**

   A critical error renders the entire content unfit for purpose or poses the risk for serious physical, financial, or reputational harm.

Calculations:

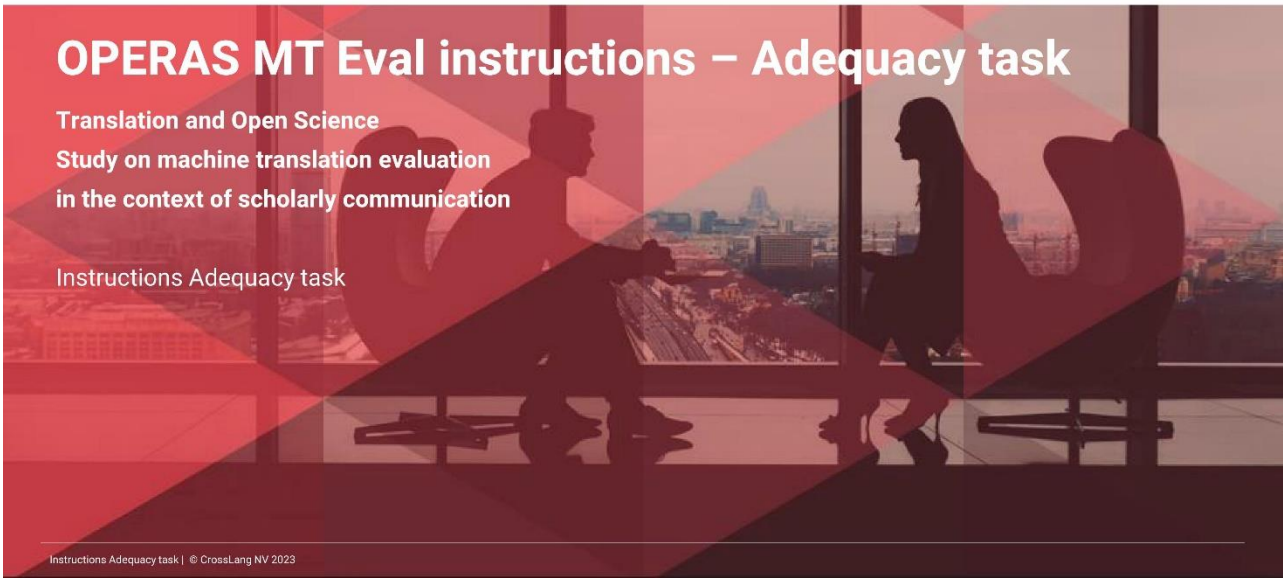
1. **Error count:** reflects the total number of instances of that individual error type or subtype assigned to a given error severity level for a given translation evaluation. The scorecard automatically multiplies the error count for each cell by its respective severity penalty multiplier to produce an intermediate product, which may or may not be displayed in a separate column in the scorecard.
2. **Absolute Penalty Total (APT):** The sum of all error type penalty totals (error counts multiplied by corresponding severity multipliers), which is used to calculate the Overall Quality Score.
3. **Per Word Penalty Total (PWPT):** PWPT is determined by dividing the absolute penalty total by the evaluation word count.

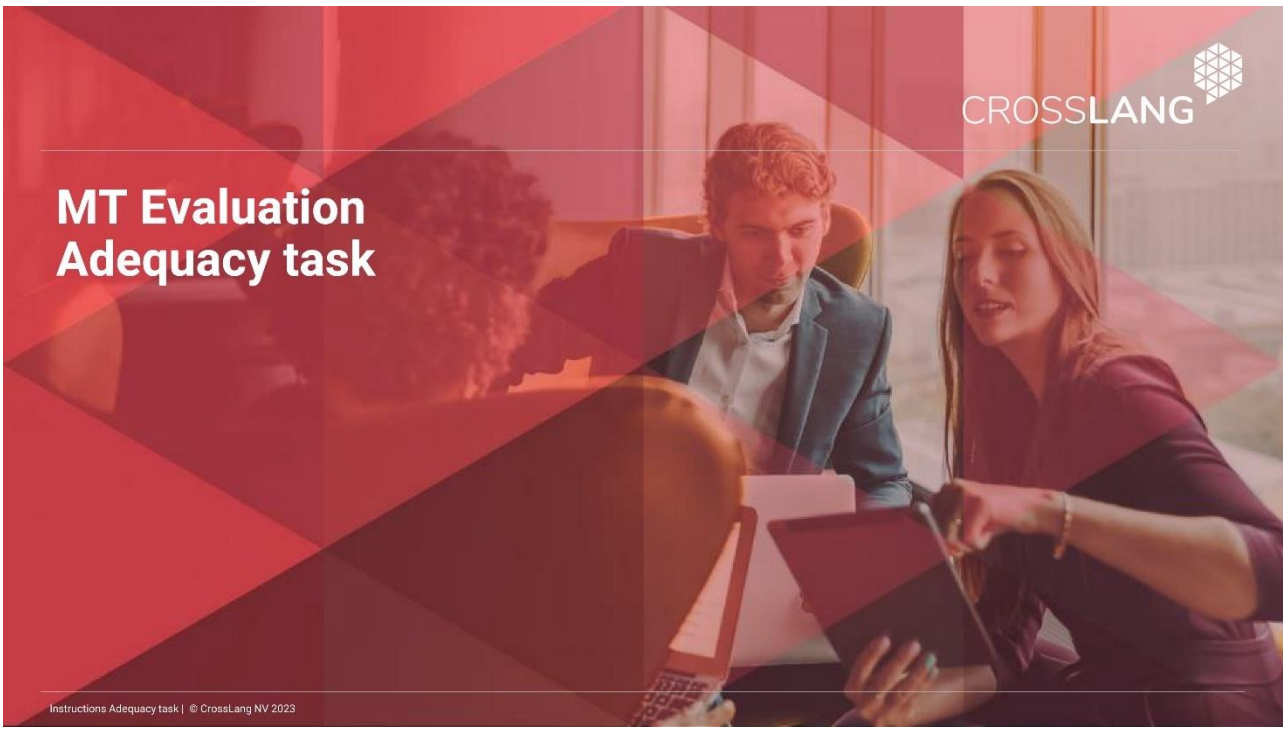
4. **Overall Quality Score (OQS):** OQS is determined by multiplying the per-word penalty score (PWPT) by the maximum score value (which is 100) and subtracting this value from 1. This process manipulates the score so that it resembles a percentage value.

	A	B	C	D	E	F	G	H
1	MQM Scorecard: Top-Level Error Typology with 4 Severity Levels							
2								
3		*Error Severity Levels:*	Neutral	Minor	Major	Critical	Error Type Penalty Total	
4		*Severity Penalty Multipliers:*	0	1	5	25		
5	ET Nos	Error Types	Error Counts				ET Weights	ETPTs
6	1	Terminology	2	7	7	0	1.0	42.0
7	2	Accuracy	4	14	7	1	1.0	74.0
8	3	Linguistic conventions	1	23	9	0	1.0	68.0
9	4	Style	5	7	3	0	1.0	22.0
10	5	Locale convention	1	12	5	0	1.0	37.0
11	6	Audience appropriateness	0	2	1	0	1.0	7.0
12	7	Design and markup	0	6	1	0	1.0	11.0
13	8	Custom						
14							Absolute Penalty Total:	261.00
15								
16		Evaluation Word Count:	10184				Per-Word Penalty Total:	0.0256
17		Reference Word Count:	1000				Overall Normed Penalty Total:	25.63
18		Scaling Parameter (SP):	1.00				Overall Quality Score:	97.44
19		Max. Score Value:	100.00					
20		Threshold Value:	85.00				Pass/Fail Rating:	Pass

Figure 6 - Sample scorecard and numerical values that represent translation quality

Annex III: Instructions for adequacy task





Evaluation Process (1)

- You will receive a notification e-mail with the link to the evaluation tool
- Click the link in the e-mail
- You will be directed to your task overview in the CrossLang evaluation tool

Notification e-mail

Hi,

A Comparison task has been assigned to you for job OPERAS Adequacy 2023 - batch 1.

Clicking the link below will take you to your task overview page.

On that page, click the Start button next to the job name to get started.

You will be presented with source and target sentence pairs, one after the other. There will be different target language versions generated with different machine translation engines.

The abstracts to which the segments to evaluate belong (or the abstracts of the articles to which they belong) will be provided to you in a separate email in order to provide you with more context. In that email, you will also be provided with a term list that may help you determine whether the proper terminology is being used.

Please read each translation and rate its quality by selecting one of the radio buttons in the Translation Quality section. Read the MT output first, then read the source text.

Evaluation criteria are as follows:

• Excellent (5): All meaning expressed in source fragment appears in the translation fragment. Your understanding is not improved by reading the ST because the MT output is satisfactory and would not need to be modified (grammatically correct / proper terminology is used / maybe not stylistically perfect but fulfills the main objective, i.e. transferring accurately all information).
• Good (4): Most meaning expressed in source fragment appears in the translation fragment. Terminology is translated in a proper way. Your understanding is not improved by reading the ST even though the MT output contains minor grammatical mistakes (word order / punctuation errors / word formation / morphology). You will not need to refer to the ST to correct these mistakes.
• Fair (3): Much meaning expressed in source fragment appears in the translation fragment. However, reading the ST helps you to better understand the MT output, because there are minor grammatical mistakes in it (word order / punctuation errors / word formation / morphology). Terminology is translated in an understandable, but not proper way.
• Poor (2): Little meaning expressed in source fragment appears in the translation fragment. Your understanding is improved considerably by reading the ST due to significant errors in the MT output (textual and syntactical coherence / textual pragmatics / morphology / terminology). You would have to re-read the ST a few times to be able to correct these errors in the MT output.
• Very poor (1): None of the meaning expressed in source fragment appears in the translation fragment. Your understanding only derives from reading the ST, is improved by reading the ST, as you could not understand the MT output. It contains serious errors in any of the categories listed above, including wrong part of speech (POS). You could only produce a translation by dismissing most of the MT output and/or re-translating from scratch.

More specific information on the criteria is provided to you in a PDF file with instructions.

Optionally, you can provide clarification or justification in the comments field.

If need be, you can return to earlier segments by clicking the Previous button.

To interrupt your review, click Pause.

To resume if you did not close your browser session, click Continue on the task overview page.

To resume if you did close your browser session, click the link in this assignment e-mail to get to the task overview page and click Continue.

The task will resume at the segment where you left off.

Please click the link below to get started:

http://mteval.crosslang.com/tasks/index/evaluator@mteval.com

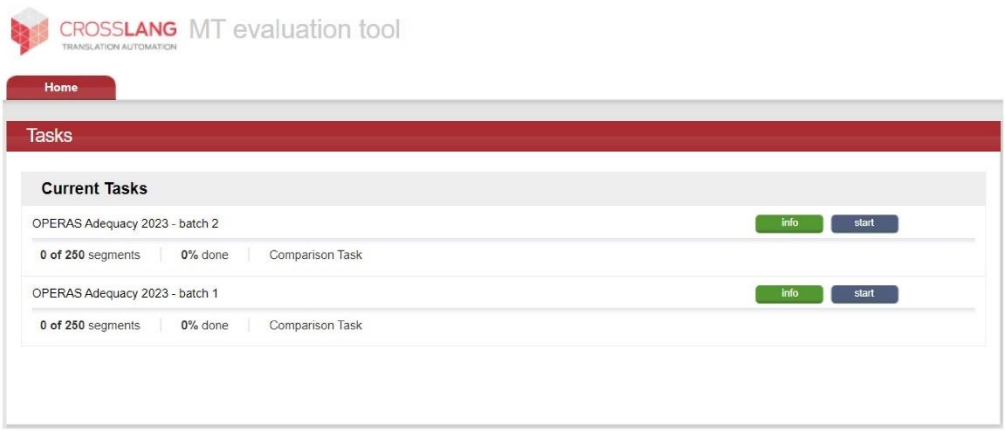
Make sure to keep this e-mail until you have completed all segments in this assignment. The link in this e-mail is the only way to get access to your task overview and continue with unfinished assignments.

Please reply to this message if you experience difficulties accessing the assigned task.

Kind regards,

The Evaluation Team

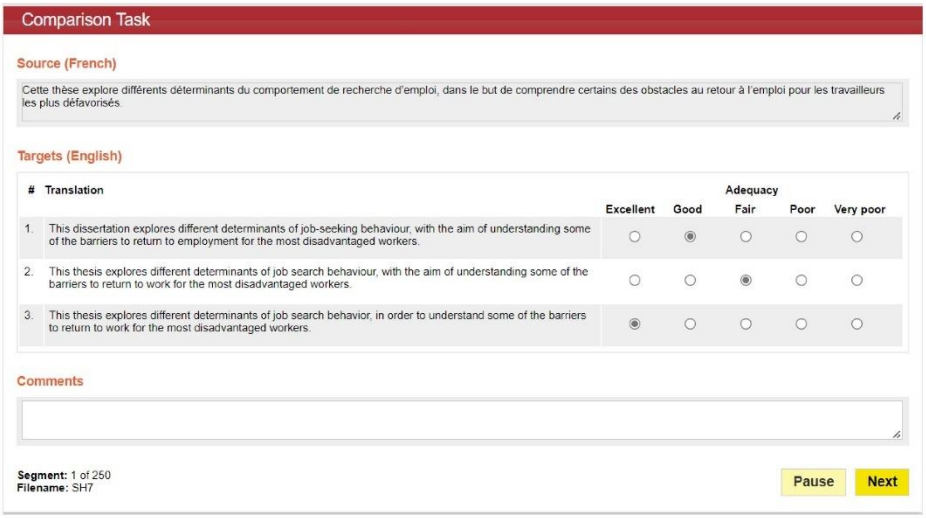
CrossLang Evaluation Tool – Task Overview



Evaluation Process (2)

- Click ‘start’ to start the evaluation task
- You will be directed to the first segment of the task

Example of a Comparison Task



Instructions Adequacy task | © CrossLang NV 2023 7

Evaluation Process (3)

- Assign a rating to each translation
- Optionally provide a comment
- Click 'Next' to move to the next segment of the task

Instructions Adequacy task | © CrossLang NV 2023 8

Adequacy Evaluation Instructions

- Human assessment of quality level per segment
- Main focus is on adequacy, but fluency and terminology should also be taken into account
- Instructions:
  - Read the machine translation (MT) output first. Then read the source text (ST)
  - Rate adequacy of translation from Excellent to Very poor
  - Optionally: Provide comments, justification, clarification

Adequacy Quality - Values

Values	Description
Excellent (5)	All meaning expressed in source fragment appears in the translation fragment. Your understanding is not improved by reading the ST because the MT output is satisfactory and would not need to be modified (grammatically correct / proper terminology is used / maybe not stylistically perfect but fulfills the main objective, i.e. transferring accurately all information).
Good (4)	Most meaning expressed in source fragment appears in the translation fragment. Terminology is translated in a proper way. Your understanding is not improved by reading the ST even though the MT output contains minor grammatical mistakes (word order / punctuation errors / word formation / morphology). You will not need to refer to the ST to correct these mistakes.
Fair (3)	Much meaning expressed in source fragment appears in the translation fragment. However, reading the ST helps you to better understand the MT output, because there are minor grammatical mistakes in it (word order / punctuation errors / word formation / morphology). Terminology is translated in an understandable, but not proper way.
Poor (2)	Little meaning expressed in source fragment appears in the translation fragment. Your understanding is improved considerably by reading the ST due to significant errors in the MT output (textual and syntactical coherence / textual pragmatics / morphology / terminology). You would have to re-read the ST a few times to be able to correct these errors in the MT output.
Very poor (1)	None of the meaning expressed in source fragment appears in the translation fragment. Your understanding only derives from reading the ST, is improved by reading the ST, as you could not understand the MT output. It contains serious errors in any of the categories listed above, including wrong part of speech (POS). You could only produce a translation by dismissing most of the MT output and/or re-translating from scratch.

Adequacy Quality – Checklist

	Excellent	Good	Fair	Poor	Very poor
Accurate transfer of meaning	All information	Most information	Much information	Little information	None of the information
Understanding by reading ST	Not improved	Not improved	Improved	Improved considerably	ST is key for understanding
Need for modification of MT output	No need	Minor changes	Minor changes	Major changes	MT output can be dismissed
Grammar	Correct	Not correct	Not correct	Not correct	Not correct
Terminology	Proper	Proper	Understandable	Not understandable	Not understandable
Style of TT	Not perfect	Not perfect	Not perfect	Not good	Not good
Type of errors		Word order/ punctuation / word formation/ morphology	Word order/ punctuation / word formation/ morphology / terminological accuracy	Textual and syntactical coherence/ textual pragmatics/ morphology / terminology	All categories of errors

Example of Terminology: Proper - Understandable - Not understandable:

Source: We estimate the reproductive age

- Nous faisons une estimation de l'âge de procréation **(proper)**

- Nous faisons une estimation de l'âge de reproduction **(understandable)**

- Nous faisons une estimation de l'âge de production **(not understandable)**

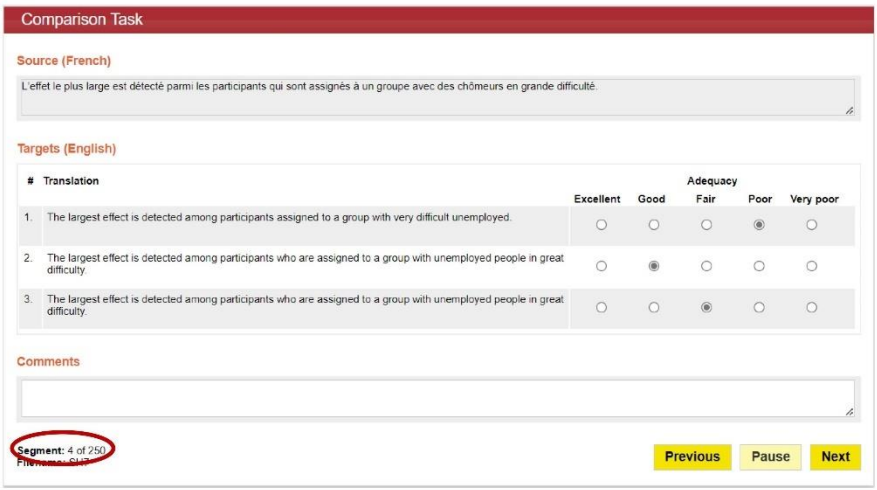
Instructions Adequacy task | © CrossLang NV 2023 11

Evaluation Process (4)

- Progress in the task is shown under each segment
- To pause your task, click 'Home' (displays the task overview) or close the web browser
- Progress in the task is also shown in the task overview
- To resume, click the button 'Continue' in the task overview
- The task will resume at the segment where you left off

Instructions Adequacy task | © CrossLang NV 2023 12

Task Progress



Instructions Adequacy task | © CrossLang NV 2023 13

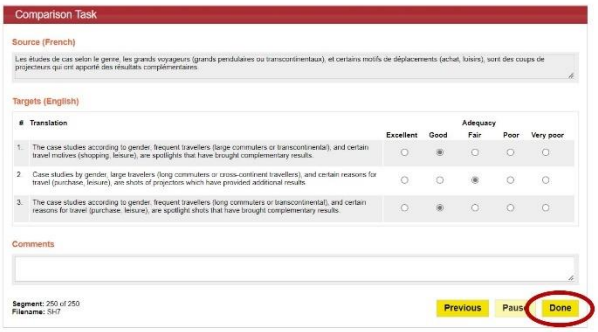
Task Progress – Task Overview

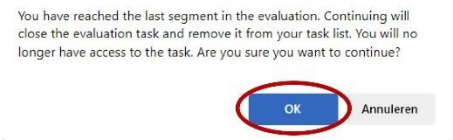
Instructions Adequacy task | © CrossLang NV 2023 14

Evaluation Process (5)

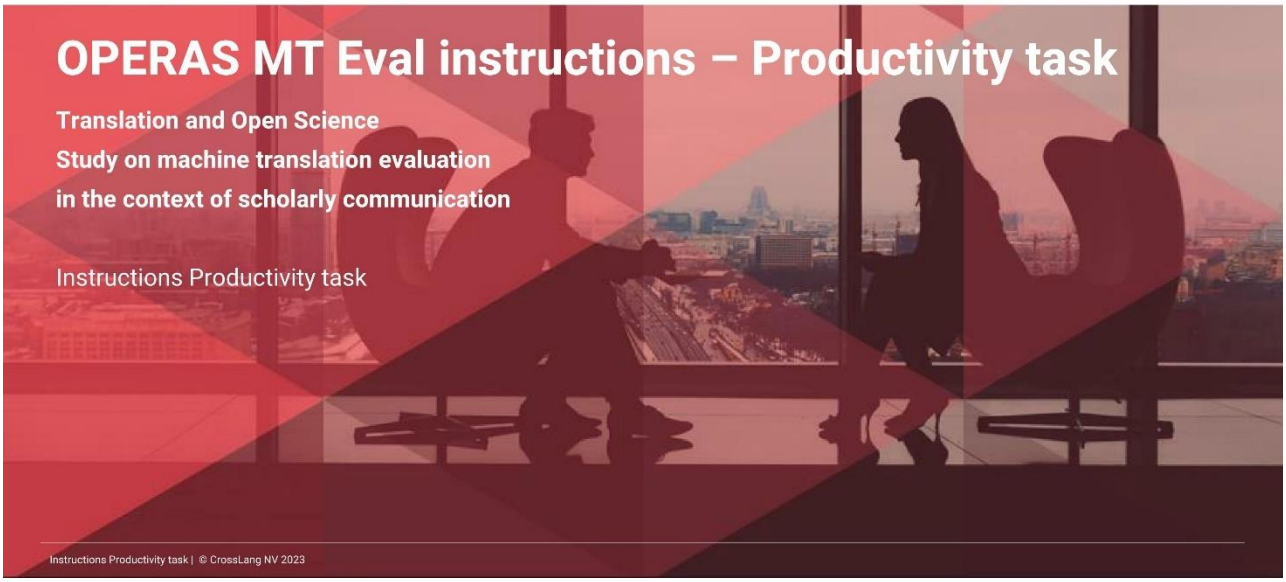
- In the last segment of the task, click ‘Done’ to finish the evaluation task
- The task will disappear from your task overview
- The CrossLang evaluation team will receive an automatic notification e-mail that the task is finished

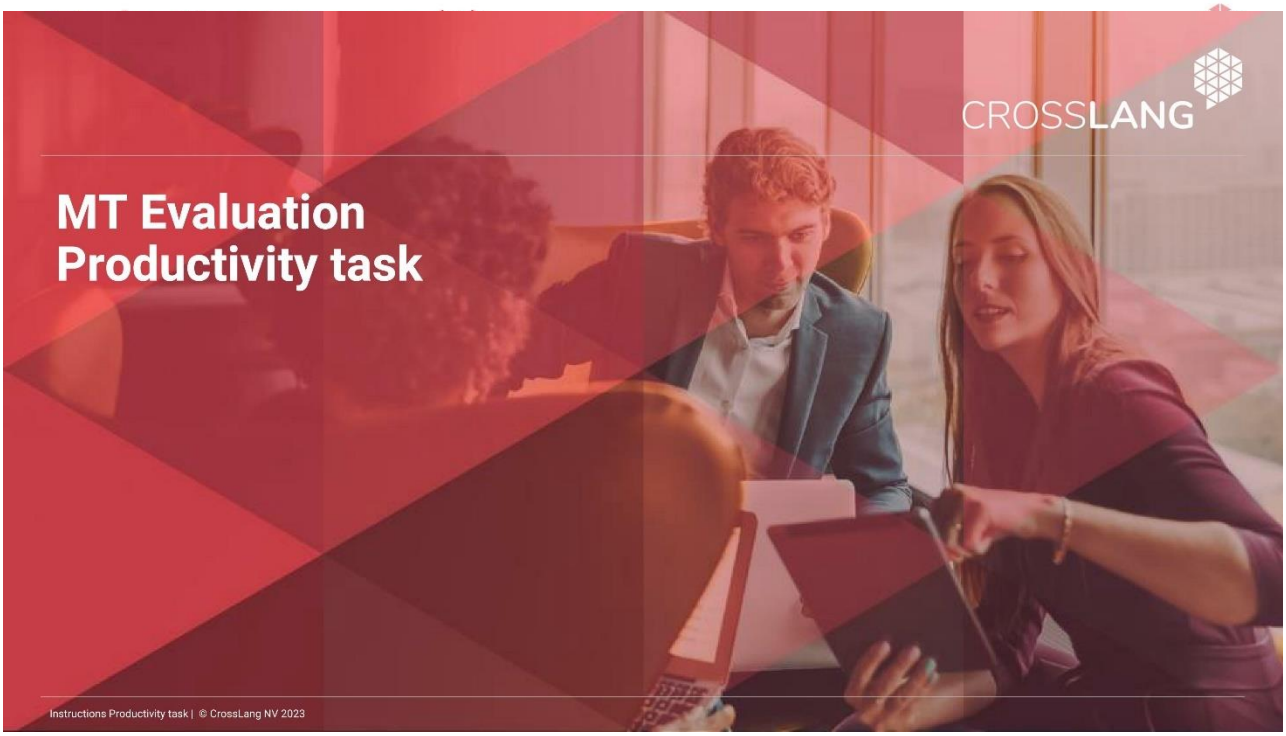
Finish Evaluation Task





Annex IV: Instructions for post-editing task





Evaluation Process (1)

- You will receive a notification e-mail with the link to the evaluation tool
- Click the link in the e-mail
- You will be directed to your task overview in the CrossLang evaluation tool

Notification e-mail

Hi,

A Productivity task has been assigned to you for job OPERAS Productivity 2023 - batch 3.

Clicking the link below will take you to your task overview page.

On that page, click on the start button next to the job name to get started.

You will see source and target sentence pairs. The context is also shown around the source segment. This includes the original MT output, preceded by the previous source segment and followed by the next source segment.

The abstracts to which the segments to evaluate belong (or the abstracts of the articles to which they belong) will be provided to you in a separate email in order to provide you with more context. In that email, you will also be provided with a term list that may help you determine whether the proper terminology is being used.

Please look at these abstracts and term list before starting the task, as we keep track of the timing needed for post-edition.

Read the MT output carefully and compare it to the source text. If the translation is correct, click Next. If the translation has mistakes, correct the mistakes and click Next.

Once you click next, you get a new view in which you can specify your perceived effort in the 'Target' box. This refers to your perception of the level of effort required to post-edit the segment. Please provide a number between 1-5:
1: Very low effort
2: Low effort
3: Moderate effort
4: High effort
5: Very high effort
More specific information is provided to you in a PDF file with instructions.

Whenever you see 'NEW PARAGRAPH' in front of the source segment, it indicates that a new paragraph will be shown. Do not take the text 'NEW PARAGRAPH' into account for the post-edition.

Note that it is not possible to go back to previously edited segments. You can only move forward to the next segment.

To interrupt post-editing, click Pause. Make sure to pause before you have made any edits. Any changes made to the current segment will be lost upon pausing.

To resume if you did not close your browser session, click Continue on the task overview page. To resume if you did close your browser session, click the link in this assignment e-mail to get to the task overview page and click Continue. The task will resume at the segment where you left off.

Please click the link below to get started:

http://mteval.crosslang.com/tasks/index/evaluator@mteval.com

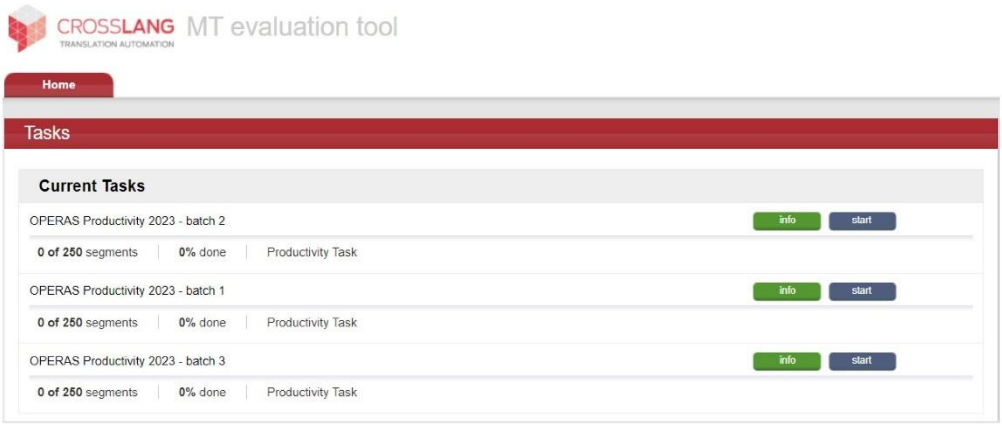
Make sure to keep this e-mail until you have completed all segments in this assignment. The link in this e-mail is the only way to get access to your task overview and continue with unfinished assignments.

Please reply to this message if you experience difficulties accessing the assigned task.

Kind regards,

The Evaluation Team

CrossLang Evaluation Tool – Task Overview



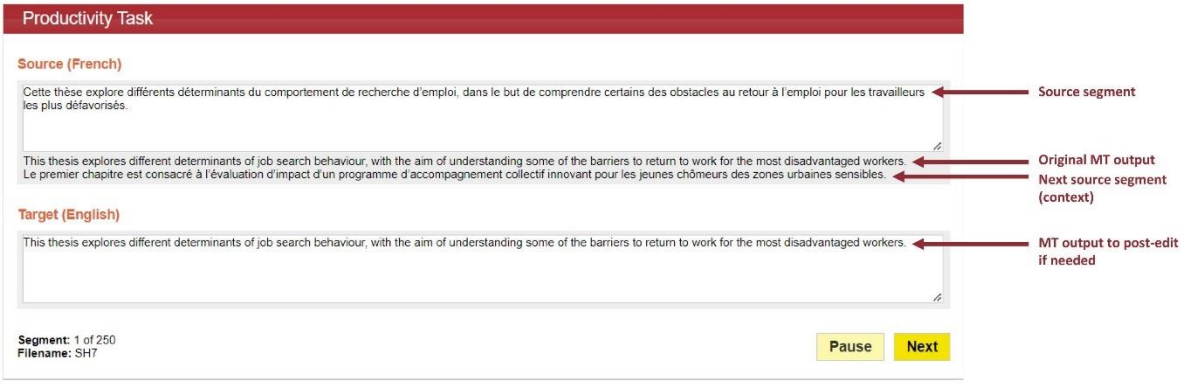
Instructions Productivity task | © CrossLang NV 2023 5

Evaluation Process (2)

- Click ‘start’ to start the evaluation task
- You will be directed to the first segment of the task

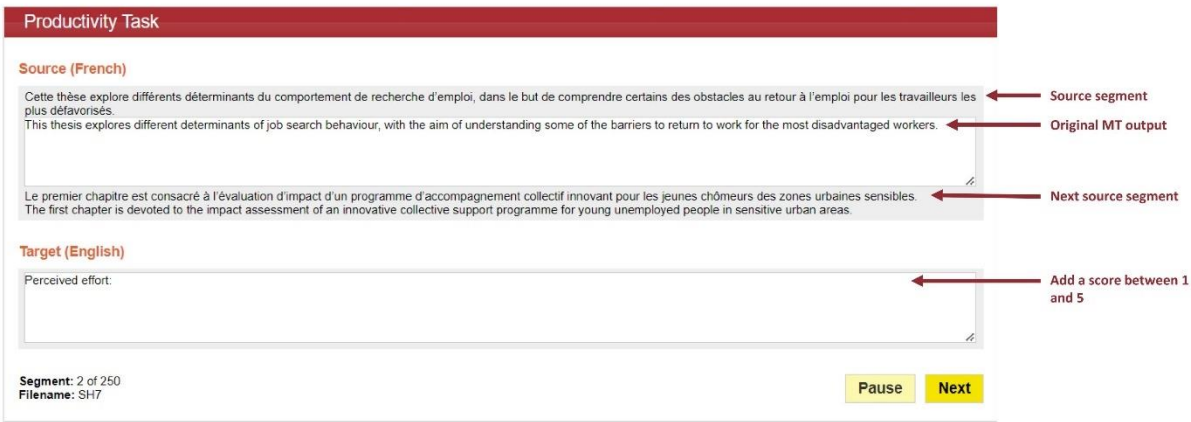
Instructions Productivity task | © CrossLang NV 2023 6

Example of a Productivity Task



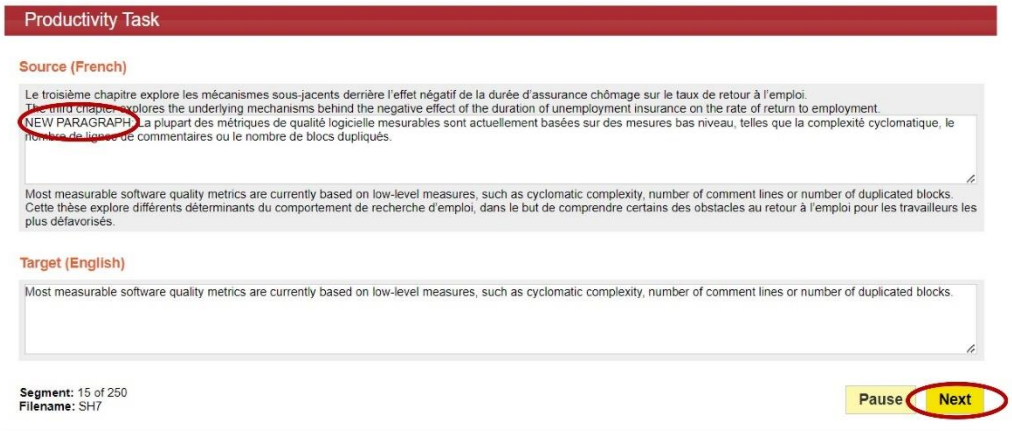
Instructions Productivity task | © CrossLang NV 2023 7

Perceived effort



CrossLang Company Presentation | © CrossLang NV 2021 8

Indication of new paragraph



This indicates that a new paragraph will be shown. Do not take the text 'NEW PARAGRAPH' into account for the post-edition.

*By 'paragraph' we mean a set of several consecutive segments from a given document (4/5 to 10 segments); these can also be titles or subtitles.
Note that paragraphs following each other may not belong to the same document.

CrossLang Company Presentation | © CrossLang NV 2021 9

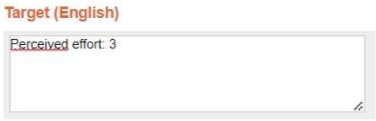
Evaluation Process (3)

- Post-edit the MT output
- Instructions:
  - If the translation is correct, do nothing.
  - If the translation has mistakes, correct them.

Instructions Productivity task | © CrossLang NV 2023 10

Evaluation Process (3)

- Specify your perceived effort:
  - Add a number on a scale of 1-5:
    - 1: Very low
    - 2: Low
    - 3: Moderate
    - 4: High
    - 5: Very high



Instructions Productivity task | © CrossLang NV 2023 11

Productivity Evaluation Instructions

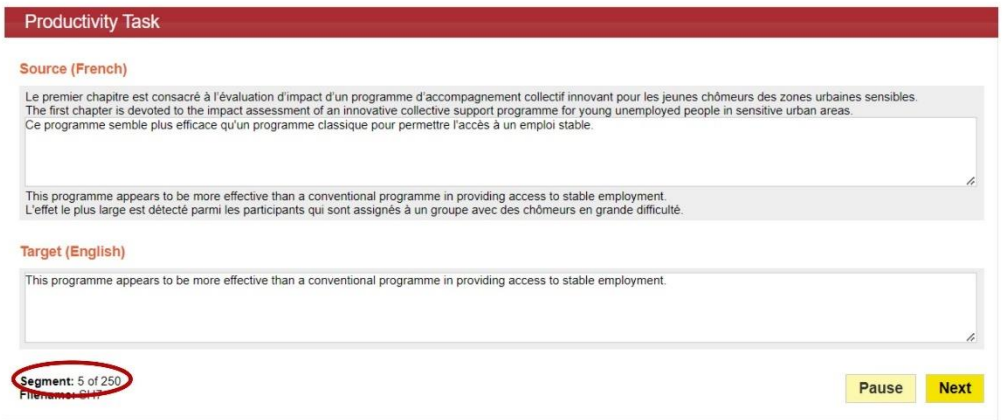
Full post-editing achieves the following objectives:
Grammar and spelling are correct
Punctuation is correct and consistent
Spelling is consistent (e.g. hyphenation)
Terminology is accurate and consistent
Style and tone are appropriate for content
Style is consistent (headers and list items)
Special formatting requirements are met (e.g. quotation marks)

Instructions Productivity task | © CrossLang NV 2023 12

Evaluation Process (4)

- Progress in the task is shown under each segment
  - Note that the number of segments is higher than that of the actual one because of the “Perceived effort” screen.
- To pause your task, click ‘Home’ (displays the task overview) or close the web browser
- Progress in the task is also shown in the task overview
- To resume, click the button ‘Continue’ in the task overview
- The task will resume at the segment where you left off

Task Progress

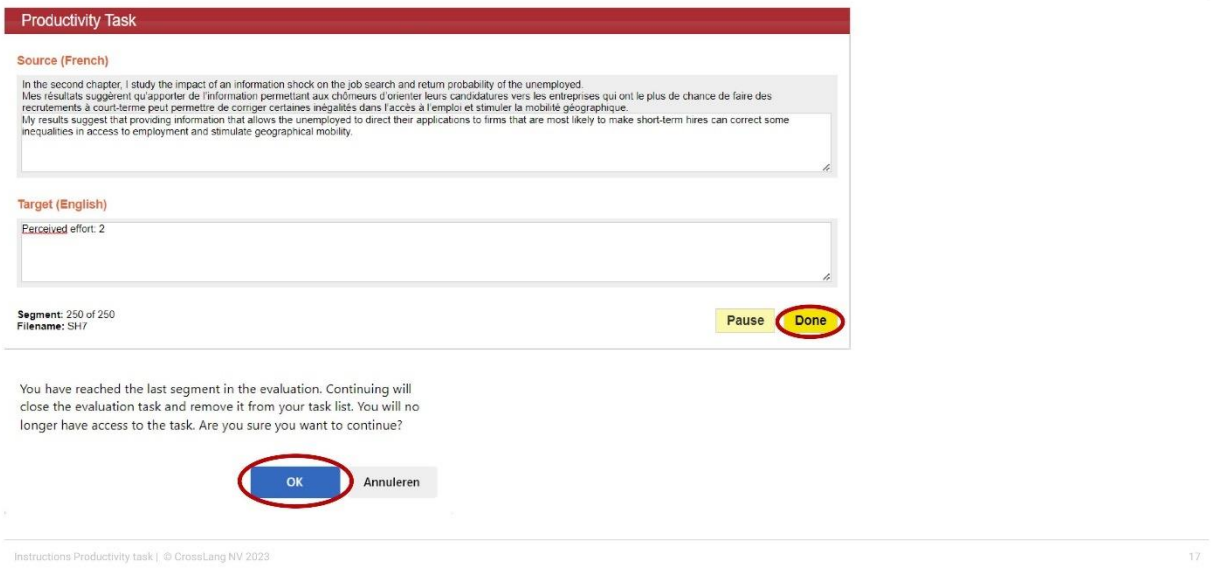


Task Progress – Task Overview

Evaluation Process (5)

- In the last segment of the task, click ‘Done’ to finish the evaluation task
- The task will disappear from your task overview
- The CrossLang evaluation team will receive an automatic notification e-mail that the task is finished

Finish Evaluation Task



Annex V: Profiles of the evaluators

Discipline 1: Human Mobility, Environment and Space

Translators

- *Translator 1:* Former Head of a Centre for Translator Training and Translation Studies in a French university, now freelance translator. Experience in a previous project of machine translation evaluation in the humanities and social sciences.
- *Translator 2:* Scientific editor and translator, postdoctoral researcher. Experience as a translator for a scientific journal specialising in environmental studies, geography, development, demography, and urban studies.

French-native researchers

- *Researcher 1:* Researcher since 2008, currently *Maître de conférences* in social geography.
- *Researcher 2:* Researcher since 2011, specializing in biodiversity.

English-native researchers

- *Researcher 3:* retired professor specializing in GIS (Geographic Information Systems), data quality testing and the social and institutional aspects of GIS.
- *Researcher 4:* PhD student since 2021, specializing in climate change.

Discipline 2: Neuroscience and Disorders of the Nervous System

Translators

- *Translator 1:* freelance translator since 2019, specializing in healthcare.
- *Translator 2:* freelance translator since 2004, specializing in neuroscience, veterinary science, social sciences and humanities.

Researchers

- *Researcher 1:* Researcher in neurosciences since 2007 in a French university. French native speaker, with perfect command of English (4 years in US and UK).
- *Researcher 2: ATER* in a French university since 2018, specializing in changes and cognitive processes underlying novel skill acquisition. Serbian native speaker, with perfect command of English and French.

Discipline 3: Climatology and Climate Change

Translators

- *Translator 1:* scientific and technical translator since 2007, specialising in environment and climate change, energy, industrial equipment and processes, electronics & optics, industrial process engineer.
- *Translator 2:* freelance translator since 2013, specialising in environment and agriculture.

Researchers

- *Researcher 1:* researcher since 2011, currently working in a French research centre for Environmental Geosciences. French native speaker, with perfect command of English.

- *Researcher 2:* researcher in Earth and Universe Sciences in a French university since 2011. Italian native speaker, with perfect command of English and French

CROSSLANG

www.crosslang.com

CrossLang NV
Amerikagebouw Kerkstraat
106 9050 Gentbrugge
Belgium

+ 32 9 335 22 00
info@crosslang.com