Data Augmentation with Translation Memories for Desktop Machine Translation Fine-tuning in 3 Language Pairs

Gokhan Dogru, Universitat Autònoma de Barcelona
Joss Moorkens, SALIS/ADAPT Centre, Dublin City University

ABSTRACT

This study aims to investigate the effect of data augmentation through translation memories for desktop machine translation (MT) fine-tuning in OPUS-CAT. It also focuses on assessing the usefulness of desktop MT for professional translators. Engines in three language pairs (English → Turkish, English → Spanish, and English → Catalan) are fine-tuned with corpora of two different sizes. The translation quality of each engine is measured through automatic evaluation metrics (BLEU, chrF2, TER and COMET) and human evaluation metrics (ranking, adequacy and fluency). Overall evaluation results indicate promising quality improvements in all three language pairs and imply that the use of desktop MT applications such as OPUS-CAT and fine-tuning MT engines with custom data in a translator's desktop can potentially provide high-quality translations aside from their advantages such as privacy, confidentiality and low use of computation power.

KEYWORDS

machine translation fine-tuning, domain adaptation, desktop machine translation, localization, parallel corpora, professional translators, machine translation evaluation.

1. Introduction

The increased quality of machine translation (MT) output since the advent of neural MT (NMT) has led to its integration into many translation and localization workflows, to the extent that MT post-editing "has now become the rule rather than exception in localization" (Esselink 2022: 90). As with other historic changes in translation production, this change has been mostly top-down, and Esselink feels that MT has been 'reluctantly' accepted. This chimes with discourse about the loss of agency (Abdallah 2012) on the part of translators when MT is unilaterally imposed rather than introduced using a participatory approach, which in turn has repercussions for translator morale and industrial sustainability. In this scenario, translators receive pre-populated MT output to post-edit, having had little or no input into the appropriateness of MT for their task and the training data used when preparing the system (Cadwell *et al.* 2018).

This article sets out an alternative scenario, in which translators themselves build their own free and open-source custom desktop NMT system to work within their familiar translation editing environment; thus, NMT becomes an empowering tool under their own control. We provide guidelines for system fine-tuning by professional translators and build on this by investigating the effects of data augmentation to ascertain the effectiveness of different amounts of data on the quality of a local NMT system. Since these NMT systems run locally, they can keep translated data secure to avoid it leaking externally, and being customizable, engines can be fine-tuned to potentially improve translation quality and consistency by adapting to the translation

https://doi.org/10.26034/cm.jostrans.2024.4716

memories (TMs) of the translators. We use OPUS-CAT (Nieminen 2021), a software collection that provides these capabilities through pretrained NMT models (Tiedemann & Thottingal 2020) and a fine-tuning feature, with plug-ins to integrate with many CAT tools including OmegaT[1], Trados[2] and memoQ[3].

The evaluation section of this article aims to measure the quality improvements, if any, in MT engines in the localization domain, fine-tuned with differently sized custom corpora. The engines are trained in English → Turkish, English → Spanish, and English → Catalan using the OPUS-CAT MT application running on the Windows operating system. While quality improvements through data augmentation are foreseeable, our aim is to show that these improvements are feasible not only in a supercomputer environment but also on a personal computer, and to explore the effect of different fine-tuning corpus sizes with the objective of guiding professional translators across various language pairs. The translation quality of each engine is measured using four automatic evaluation metrics, namely BLEU (Papineni et al., 2002), chrF2 (Popović 2015), TER (Snover *et al.* 2006) and COMET (Rei *et al.* 2020), along with human evaluation metrics (ranking, adequacy and fluency).

Pretrained NMT models from OPUS-MT are mostly trained on mixed domain corpora, therefore specific TMs need to be added to adapt the style and terminology to different specific domains. However, empirical studies on how large domain-specific TMs should be to provide significant quality improvements in the relevant domain are needed. Our study concentrates on the localization domain in three language pairs and measures translation quality in three scenarios: i) no fine-tuning, ii) fine-tuning with a bilingual localization corpus of 500,000 source words and iii) fine-tuning with a bilingual localization corpus of more than 2,000,000 source words. Evaluating the quality of each engine with both automatic and human evaluation metrics allows us to observe how adding custom parallel corpora affects MT translation quality.

Translation corpora are obtained from Microsoft Visual Studio's *Translation and UI Strings*[4] for the English → Turkish and Spanish → Turkish language pairs, and from SoftCatalà for the English → Catalan language pair. These are compiled as TMs to be used as fine-tuning corpora. A total of 210 sentences in the localization domain are selected for automatic and human evaluation tasks. Human evaluation is conducted using three metrics (adequacy, fluency and ranking) within the KantanLQR[5] platform by three reviewers per language pair. KantanLQR allows for customizing quality evaluation metrics to be used for evaluation and provides an interface for evaluation tasks to be streamlined together with a dashboard for a quick overview of the results.

While we expect to observe quality improvements with each additional localization corpus, fine-tuning does not necessarily guarantee such an improvement. Our findings will provide insights for translators who would like to build and manage their own secure MT systems, effectively augmenting MT with their domain-appropriate data. It should be noted that usefulness in the context of this study is taken from a broader perspective, seeing MT as a resource in the workflow of the translator, not necessarily concerned with productivity gains through higher quality MT engines, but also highlighting tertiary issues such as control over data, transparency, and confidentiality. Nonetheless, improvements in the MT performance following fine-tuning steps by professional translators may also imply more usefulness.

2. Related Work

Research on translator interaction with NMT has tended to focus on productivity or quality rather than its "usefulness… as a tool for professionals", focusing instead on improving the NMT systems themselves (Ragni & Nunes Vieira 2022: 153). Research on human factors in MT, for example, tends to focus on post-editing effort and productivity, although measurement of keystrokes or their approximation using the Human-targeted Translation Edit Rate (HTER; Snover *et al.* 2006) metric gives an indication of the usefulness of MT. Studies on user interfaces (UIs) for translator interaction with MT aim to make MT more useful so that interactions become more user friendly with reduced cognitive friction (e.g. Moorkens and O'Brien 2017; Herbig *et al.* 2020). However, the usefulness of MT is again not the focus of such research.

Studies such as those of Kenny and Doherty (2014), Martín Mor (2017), Ramírez-Sánchez *et al.* (2021) and Kenny (2022) have highlighted the didactics of teaching MT to translators. Free and open-source platforms such as MTradumàtica[6] (for statistical MT) and MutNMT[7] (for NMT) have allowed translators to experiment with all steps of MT training in an experimental environment. The availability of these platforms helps professional translators understand the capabilities and limitations of MT, and make informed decisions about their uses of MT. These platforms are built for educational purposes, for students and professional translators who would like to integrate MT into their workflow using a stable, easy-to-use and flexible tool.

The convergence of different projects within the OPUS platform (Tiedemann *et al.* 2022) such as OPUS Corpus (Tiedemann 2012), OPUS-MT (Tiedemann and Thottingal 2020) and OPUS-CAT (Nieminen 2021) has, among other things, paved the way for translators to use MT in different ways in their workflow. The release of OPUS-CAT has particularly bridged the gap between MT research and professional use of MT by translators. OPUS-CAT is a software collection with a graphical UI that runs on Windows; it allows translators to use pretrained NMT models from OPUS-MT and fine-tune them with their TMs (or TMs from their clients or other sources of free and open-source corpora) and connect them into their CAT tool environment. Such a setup has many advantages. For example, it lets the translator assume control of the MT system and to regularly update MT engines with

TMs without allowing client data to leak to third parties. Furthermore, the presence of pretrained NMT models decreases computational costs and allows for reuse of these models, which reduces environmental and energy cost. This setup helps to solve some concerns related to transparency, confidentiality, unethical data use, and privacy as highlighted in Moorkens and Lewis (2019) and Moorkens (2022).

Finally, localization (Esselink 2003) has been one of the fastest-growing domains in the language industry. However, there are few academic studies on the domain in general (Jiménez-Crespo 2020; Ramos *et al.* 2022). It is particularly hard to find studies that focus on the use of MT in localization scenarios. The study herein aims to provide baseline results from different language pairs on MT and localization, a domain which is characterized by inline format tags, variables, adaptation aspects, short strings, and context dependencies, all of which are known to cause problems for MT.

3. Methodology and Research Design

Three types of MT engines were used or created per language pair. The first type of engine is a pretrained model from OPUS-MT (Tiedemann and Thottingal 2020). It was downloaded to OPUS-CAT (Nieminen 2021) through the built-in feature "Install OPUS Model from Web". Once the download was complete the engine was ready for translation. This type of engine is referred to as the "baseline model" throughout the present study. The pretrained models have the advantage of not requiring the end user to train an engine from scratch. This means that translators do not need to spend huge amounts of money on expensive hardware for NMT training or for electricity for resource-intensive computation during training. Once trained, pretrained NMT engines can be used and shared without the need to repeat this process, making them more environmentally friendly and sustainable (Tiedemann *et al.* 2022:1).

English → Turkish[8], English → Spanish[9] and English → Catalan[10] baseline pretrained NMT models are hosted in the GitHub repository of the Language Technology Research Group at the University of Helsinki. The second type of engine was created by fine-tuning these baseline models with a localization corpus of approximately 500,000 source words extracted randomly from the larger versions of the corpora. Finally, the third type of engine was created by fine-tuning the baseline model with a localization corpus that has between 2,300,000 and 2,700,000 source words. The exact source and size of each corpus is described in Section 3.1.

Fine-tuning was conducted within OPUS-CAT by selecting the baseline model ("Fine-tune selected model"), importing the relevant TMX file and providing a specific name to the prospective fine-tuned engine in the next window and clicking the "Fine-tune" button. With the default fine-tuning settings (a single thread and a workspace of 2048 MB; stopping after one epoch; learning rate: 0.00002), the training time varies and can last for long durations depending on the size of the fine-tuning corpus and the computational power used. In our study, we use a laptop with 16 GB RAM, GEForce MX150 graphic card (total available graphic memory: 10183 MB), and Intel Core i7-8550 CPU processor. With these specifications and default fine-tuning settings in OPUS-CAT, it takes approximately 4 hours to fine-tune with 500,000 source words and approximately 10 hours with 2,000,000 source words. It is possible to change fine-tuning parameters such as epochs and learning rate. For this study, we kept the default settings, assuming that a translator using OPUS-CAT would not make any change to these parameters.

3.1. Corpus Statistics

Aside from the baseline model that does not include additional fine-tuning, the study involves fine-tuning pre-trained engines with two different localization corpus sizes: 500,000 source words and more than 2,000,000 source words. Parallel corpora in English → Turkish, English → Spanish and English → Catalan were compiled from resources available on the web and used in the TMX format. English → Turkish and English → Spanish corpora were obtained from Microsoft Visual Studio's *Translation and UI Strings*. These corpora are available as sets of various CSV (Comma Separated Values) files. The files were consolidated into a single TMX file using memoQ's multilingual delimited text filter which allows conversion of a bilingual spreadsheet into TMX in a few steps. The English → Turkish corpus includes 2,300,000 source words while English → Spanish includes 2,700,000 source words. Corpora sizes across language pairs differ since the original source files in Visual Studio are of different sizes depending on the language pair. The study aimed to use all available corpora to the extent possible. While we tried to keep corpora sizes similar, it was not necessary for them to be the same since the main objective of the study is not to make comparisons across language pairs but focuses on quality improvements through data augmentation. Hence, different sizes in the large corpus scenario may provide different insights for professional translators.

The English → Catalan corpus in Microsoft Visual Studio, containing less than 800,000 source words, was deemed too small for fine-tuning in this language pair and therefore not utilized. Instead, TMs from Softcatalà[11], a nonprofit association that localizes free and open-source applications into Catalan, were compiled as a single TMX file to yield a larger corpus. Localization projects realized by this initiative include Mozilla, Bitcoin, Libre Office, Ubuntu, WordPress among others. The resulting TMX file used in the present study has 2,300,000 source words.

Out of these three large corpora, approximately 500,000 source words were copied and saved as separate, smaller corpora. These smaller corpora were used for fine-tuning the pretrained engines. Subsequently, larger versions of the TMX files were used to fine-tune the baseline model. Table 1 provides the detailed statistics of each engine, corpora sources, and engine names.

Language Pair	***Baseline***	***Small Corpus***	***Large Corpus***	***Corpus Source***
EN-TR	No fine-tuning (en-tr-1)	501,371 (en-tr-2)	2,253,304 (en-tr-3)	Microsoft
EN-ES	No fine-tuning (en-es-1)	501,979 (en-es-2)	2,745,645 (en-es-3)	Microsoft
EN-CA	No fine-tuning (en-ca-1)	503,188 (en-ca-2)	2,269,533 (en-ca-3)	Softcatalà

Table 1. Corpus statistics for fine-tuning each engine.

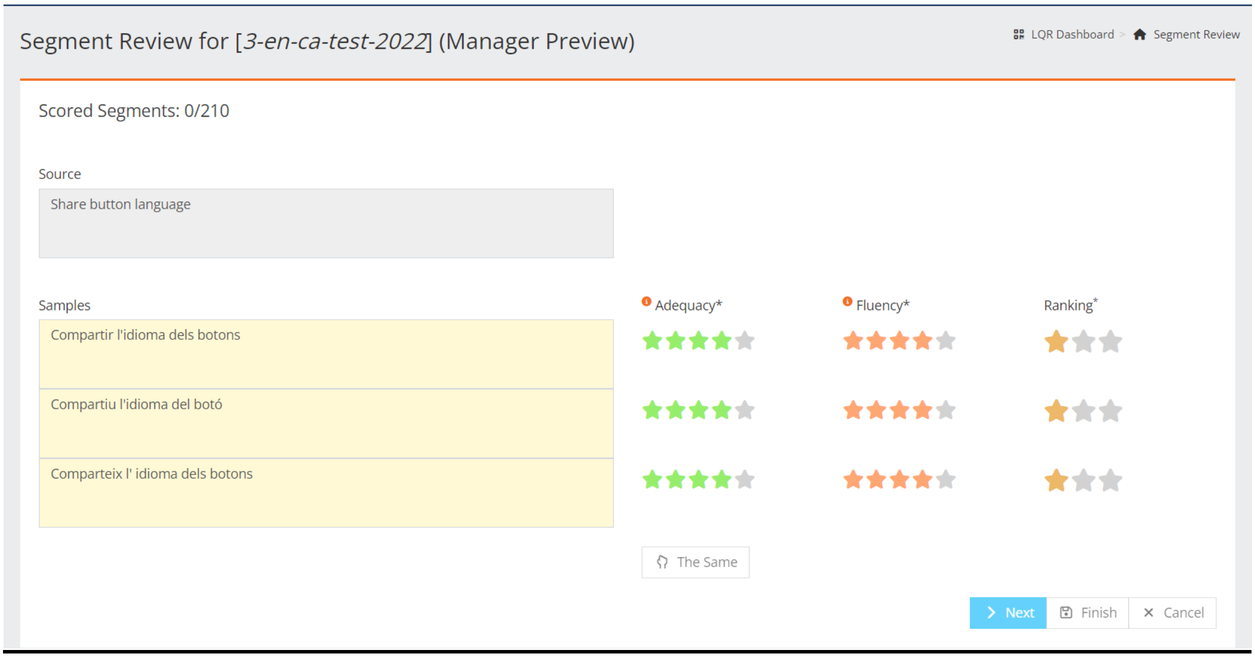
One file from the Microsoft corpus was not used for fine-tuning and was instead allocated for automatic and human evaluation. This file was present in the three target languages; hence we were able to use the same file for the evaluation tasks. In each target file, any segments that exist in the large corpus for fine-tuning or repetitions within the file were omitted and only unique segments were left. While most of the source segments were the same across three language pairs, the segment omitting steps led to slight changes. Hence, the 210 segments are not exactly the same across language pairs. These 210 segments were then selected for evaluation for each language. Corpora for all language pairs are available in GitHub[12] (together with evaluation test set and evaluation results).

3.2. MT Evaluation

Both automatic and human evaluation were employed to evaluate quality. BLEU, chrF2, TER and COMET were the automatic evaluation metrics used with human reference translations. The automatic evaluation was completed using the MATEO[13] platform (Vanroy *et al.* 2023) by uploading sample MT outputs and human translations one by one. MATEO has the advantage of providing confidence intervals and p-values for detecting significant differences between baseline engines and fine-tuned engines. Once the automatic evaluation was complete, human evaluation by professional translators was initiated.

Three professional translators per language pair participated in the evaluation tasks. All reviewers are native speakers of the target language and have extensive experience in the translation industry. Four reviewers reported more than 10 years of experience, two of them have 5–10 years of experience, two of them have 3–5 years, while one reviewer has 1–2 years of experience. All instructions (see Annex I) for the evaluation task were sent to the reviewers via email and a complete list of instructions about the evaluation platform was provided. They completed the three evaluations (ranking, adequacy, and fluency) together, one segment at a time.

The translators were asked to rank the three MT outputs from the best to the worst by assigning three points to the highest-performing engine and one point to the worst performing. The order in which MT outputs was shown was randomized to avoid biases towards any engine. Once a rating was completed for a segment, the translators moved to the next window to rate the following segment. If MT outputs were considered to be of identical quality for two or more engines, equal scores were permitted.


Segment Review for [*3-en-ca-test-2022*] (Manager Preview)

Scored Segments: 0/210

Source

Share button language

Samples

Compartir l'idioma dels botons

Compartiu l'idioma del botó

Comparteix l' idioma dels botons

Figure 1. A snapshot of the evaluation screen.

Then, translators rated the adequacy and fluency of the output using a scale of five, where five was the highest score and one the lowest. The tasks were completed within the interface of the KantanLQR platform. The interface showed one source segment and three MT outputs as well as ranking, adequacy, and fluency rating options, as may be seen in Figure 1. Table 2 shows the definitions and rating scales for adequacy and fluency according to KantanLQR. Reviewers had access to this information each time their mouse hovered over the "i" icon next to adequacy and fluency.

Adequacy	**Fluency**
Adequacy measures how much meaning is expressed in the machine translation segment. It is measuring whether the machine translation segment contains as much of the information as a human translation.	Fluency is checking that the translation follows common grammatical rules and contains expected word collocation. This category scores whether the machine translation segment is formed in the same way a human translation would be
1- None of the meaning expressed in the source fragment is expressed in the translation fragment. 2- Little of the source fragment meaning is expressed in the translation fragment. 3- Much of the source fragment meaning is expressed in the translation fragment. 4- Most of the source fragment meaning is expressed in the translation fragment. 5- All meaning expressed in the source fragment appears in the translation fragment	1- No fluency. Absolutely ungrammatical and for the most part doesn't make any sense. Translation has to be rewritten from scratch. 2- Little fluency. Wrong word choice, poor grammar and syntactic structure. A lot of post-editing required. 3- Quite fluent. About half of the translation contains errors and requires post-editing. 4- Near native fluency. Few terminology or grammar errors which don't impact the overall understanding of the meaning. Little post-editing required 5- Native language fluency. No grammar errors, good word choice and syntactic structure. No post-editing required.

Table 2. Adequacy and fluency rating scale on KantanLQR.

4. Results

This section includes automatic and human evaluation results. Firstly, overall evaluation results are presented, and then a breakdown is reported per language pair.

4.1. Automatic Evaluation Results

Automatic evaluation results show how the performance of each engine differs according to BLEU, chrF, TER and COMET metrics when either a small or large corpus is used for fine-tuning. Table 3 provides the results of the automatic evaluation for each language pair.

System 1 is the baseline, System 2 fine-tuned with a small corpus added, and System 3 with the larger corpus. As indicated in MATEO, p-values show the significance of the difference between a system and the baseline. The platform puts an asterisk * to indicate that a system differs significantly from the baseline model ($p<0.05$) and best system per metric in the language pair is highlighted in bold. We use the same format.

system	***comet (μ ± 95% CI)***	***BLEU (μ ± 95% CI)***	***chrF2 (μ ± 95% CI)***	***TER (μ ± 95% CI)***
Baseline: en-tr-1	84.0 (84.0 ± 2.0)	23.0 (23.0 ± 4.9)	54.0 (54.1 ± 3.6)	65.8 (65.8 ± 5.6)
en-tr-2	89.6 (89.6 ± 1.5) (p = 0.0010)*	49.6 (49.8 ± 5.1) (p = 0.0010)*	**68.1 (68.2 ± 3.4) (p = 0.0010)***	44.2 (44.1 ± 4.9) (p = 0.0010)*
en-tr-3	**90.5 (90.6 ± 1.3) (p = 0.0010)***	**51.7 (51.8 ± 5.1) (p = 0.0010)***	67.8 (67.8 ± 3.5) (p = 0.0010)*	**42.8 (42.7 ± 5.0) (p = 0.0010)***
Baseline: en-es-1	85.0 (85.1 ± 2.0)	37.3 (37.4 ± 4.3)	66.6 (66.7 ± 2.9)	46.1 (46.0 ± 4.7)
en-es-2	87.6 (87.7 ± 1.8) (p = 0.0010)*	38.5 (38.8 ± 6.4) (p = 0.2747)	70.0 (70.1 ± 2.9) (p = 0.0100)*	39.3 (39.1 ± 3.5) (p = 0.0010)*
en-es-3	**89.6 (89.7 ± 1.5) (p = 0.0010)***	**48.1 (48.2 ± 4.3) (p = 0.0010)***	**74.6 (74.7 ± 2.5) (p = 0.0010)***	**34.9 (34.8 ± 3.6) (p = 0.0010)***
Baseline: en-ca-1	84.3 (84.2 ± 2.0)	38.0 (37.9 ± 4.8)	63.2 (63.2 ± 3.5)	57.9 (57.9 ± 6.1)
en-ca-2	84.6 (84.6 ± 2.1) (p = 0.2178)	42.5 (42.3 ± 5.3) (p = 0.0320)*	63.3 (63.3 ± 3.7) (p = 0.3836)	49.0 (48.9 ± 6.5) (p = 0.0010)*
en-ca-3	**86.6 (86.6 ± 2.0) (p = 0.0010)***	**47.2 (47.0 ± 5.1) (p = 0.0010)***	**67.8 (67.8 ± 3.3) (p = 0.0040)***	**44.0 (43.9 ± 5.7) (p = 0.0010)***

Table 3. Automatic evaluation results.

In the following three subsections, we report the results for each language pair.

4.1.1. English → Turkish MT Engines

The baseline English → Turkish MT engine has the lowest BLEU score of the nine engines in the study. However, when it was fine-tuned with the smaller localization corpus (en-tr-2), the BLEU score improved considerably from 23 to 49.6. When the large corpus was used for fine-tuning (en-tr-3), the score increased further. However, as may be observed from Table 3, although the size of the custom corpus is larger, the improvement from en-tr-2 to en-tr-3 remains modest. Similarly, in the case of chrF2, the score improves considerably when either a small or large corpus is introduced for fine-tuning. However, en-tr-2 has only a slightly higher (less than one point) score than en-tr-3 (68.1 vs 67.8, respectively). Measuring the fewest possible editing steps from MT to the human reference with TER so that a lower score implies better quality output, small and large corpora improved scores considerably. Akin to the BLEU scenario, fine-tuning with the large corpus improved the score slightly compared to fine-tuning with the small corpus (44.2 vs. 42.8, respectively). COMET scores also imply a gradual improvement with the addition of in-domain corpora. Nevertheless, as can be inferred from Table 4, the difference between en-tr-2 and en-tr-3 is only significant in COMET score and no significant change is observed in the other three metrics.

system	***comet (μ ± 95% CI)***	***BLEU (μ ± 95% CI)***	***chrF2 (μ ± 95% CI)***	***TER (μ ± 95% CI)***
en-tr-2	89.6 (89.6 ± 1.5)	49.6 (49.8 ± 5.1)	**68.1 (68.2 ± 3.4)**	44.2 (44.1 ± 4.9)
en-tr-3	**90.5 (90.6 ± 1.3)** **(p = 0.0480)***	**51.7 (51.8 ± 5.1)** **(p = 0.1079)**	67.8 (67.8 ± 3.5) (p = 0.2837)	**42.8 (42.7 ± 5.0)** **(p = 0.1319)**

Table 4. A comparison of the en-tr-2 to en-tr-3 MT systems in terms of automatic evaluation metrics. This comparison shows the impact of increasing the fine-tuning corpus from approx. 500,000 words to approx. 2,000,000 source words in this language pair.

These overall scores imply that for English → Turkish engines, even a small fine-tuning corpus can improve translation quality considerably, while the effect of a much larger fine-tuning corpus may only improve quality marginally when compared to the small corpus. Figure 2 provides a graphical depiction using all four evaluation metrics.

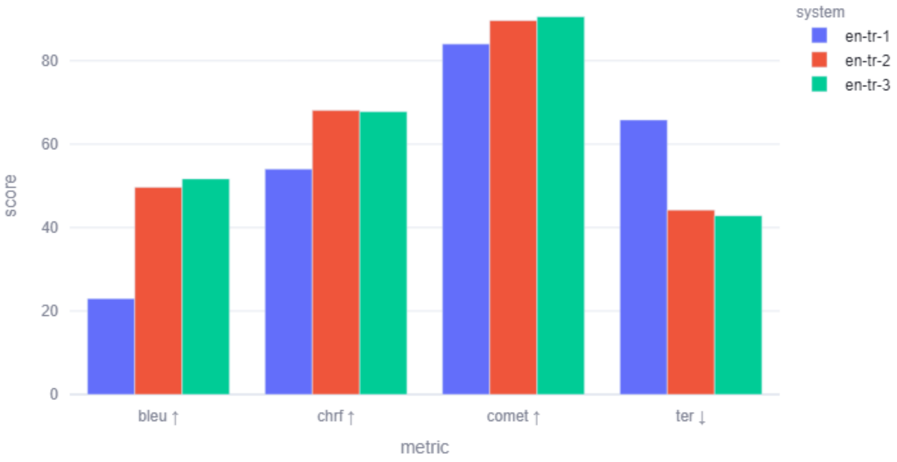


Figure 2. Comparison of the automatic evaluation scores in English → Turkish.

The MATEO platform also provides COMET scores per sentence and exports all compared sentences together with their individual scores into a spreadsheet. This possibility opens the way for more fine-grained qualitative and quantitative analysis of the outputs. Using this spreadsheet, we filtered the 51 sentences that include tags or placeholders (which are key parts of localization projects) to create an overview of how they are handled in each engine. The average COMET scores for these 51 sentences per engine are as follows: 72.3, 90 and 91.1. A closer look at the segments show that the en-tr-1 engine usually totally or partially omits tags or placeholders, or literally translates them. En-tr-2 tends to (correctly) keep the tags untranslated, but is still inconsistent and sometimes translates tags and placeholders. Finally, the en-tr-3 engine keeps tags and placeholders untranslated, retaining symbols ([], {}, / etc.) correctly. See Table 5 for a few examples. This analysis shows that the further addition of large corpus in the localization domain can provide better handling of tags and placeholders.

source	reference	en-tr-1	en-tr-2	en-tr-3
Click to \{swiftAction}	\{swiftAction} düğmesine tıklayın	\ swiftAction} 'a tıklayın	\{swiftAction} öğesine tıklayın	\swiftAction} öğesine tıklayın
Avatar of \{title}	\{title} avatarı	Avatar \ [başlık]	\{title} için Avatar	\{title} Avatarı
Back to \{section}	\{section} bölümüne dön	\} bölüme geri dönelim.	\{bölüm}'e geri dön	\{section} durumuna geri dön
Close \{topic}'s profile.	\{topic} profilini kapatın.	Profili kapat.	\{topic} profilini kapatın.	\{topic} profilini kapatın.
Welcome back, \{displayName}	Tekrar hoş geldiniz \{displayName}	Tekrar hoş geldiniz.	Tekrar hoş geldiniz, \{displayName}	Tekrar hoş geldiniz, \{displayName}

Table 5. A selection of the sentences with tags or placeholders translated by three different English → Turkish systems.

4.1.2. English → Spanish MT Engines

All automatic evaluation metrics scores rise when localization corpora were introduced for fine-tuning in the English → Spanish language pair. However, unlike the English → Turkish engines, the addition of the small corpus did not lead to a significantly improved BLEU score. Yet, when the large corpus was used for fine-tuning, the score improved considerably in all four metrics. The BLEU score is 37.3 using the baseline engine while it is 38.5 for en-es-2 and 48.1 for en-es-3. Using chrF2, the baseline engine scored 66.6, en-es-2 70.0 and en-es-3 74.6. Using TER, the baseline engine has a score of 46.64 while it improves to 43.10 for the small corpus and to 38.12 in the large corpus scenario. COMET scores also suggest significant improvements when further corpora are added when compared to the baseline.

system	*comet (μ ± 95% CI)*	*BLEU (μ ± 95% CI)*	*chrF2 (μ ± 95% CI)*	*TER (μ ± 95% CI)*
en-es-2	87.6 (87.7 ± 1.8)	38.5 (38.8 ± 6.4)	70.0 (70.1 ± 2.9)	39.3 (39.1 ± 3.5)
en-es-3	**89.6 (89.7 ± 1.5)** **(p = 0.0010)***	**48.1 (48.2 ± 4.3)** **(p = 0.0010)***	**74.6 (74.7 ± 2.5)** **(p = 0.0010)***	**34.9 (34.8 ± 3.6)** **(p = 0.0020)***

Table 6. A comparison of the en-es-2 to en-es-3 MT systems in terms of automatic evaluation metrics. This comparison shows the impact of increasing the fine-tuning corpus from approx. 500,000 words to approx. 2,000,000 in this language pair.

As may be seen in Table 6 and Figure 3, automatic comparison of the en-es-2 engine to the en-es-3 engine shows that the large localization corpus fine-tuning has brought significant and considerable improvements across all metrics.

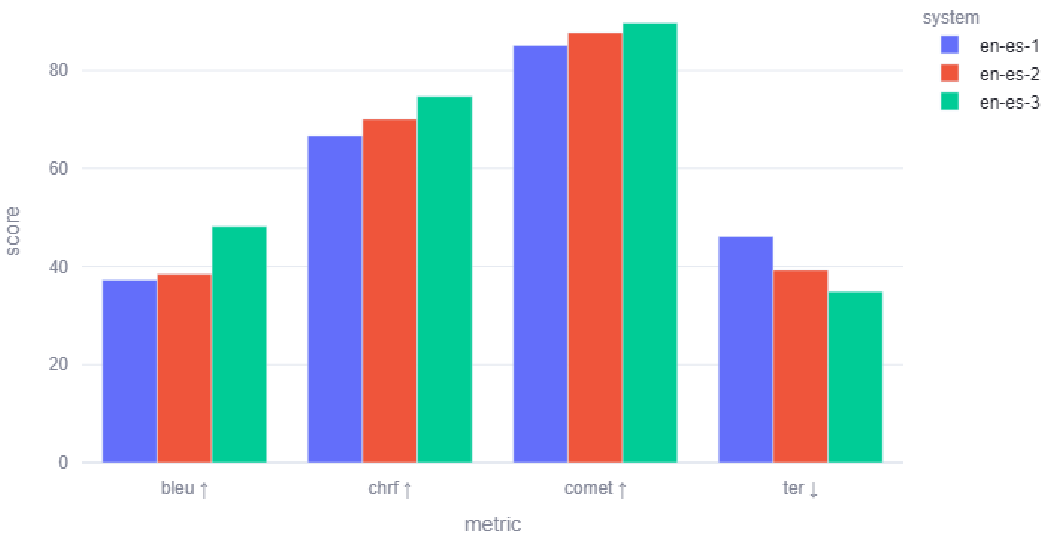


Figure 3. Comparison of the automatic evaluation scores in English → Spanish.

As with the English → Turkish case, we filter the 52 sentences with tags and placeholders to analyze the behavior of the engines. Average COMET scores for each engine for these sentences are as follows: 71, 80 and 83. The baseline en-es-1 engine does not seem to keep the tags or placeholders correctly with full or partial omissions or the introduction of different symbols such as "#". While en-es-2 and en-es-3 are more consistent with treatment of tags and placeholders, they do not conserve the form of the tags or placeholders, often converting the opening "{" into "\". Even fine-tuning with a large localization corpus does not help to solve this issue in this language pair.

source	**reference**	**en-es-1**	**en-es-2**	**en-es-3**
Voted by (\{count})	\{count} votos	Votado por ('cuento})	Votado por (######count})	Votado por (\\count})
\{participantName}. More options.	\{participantName}. Más opciones.	â € ¢participantName}. Más opciones.	\\participantName}. Más opciones.	\\participantName}. Más opciones.
Already using Skype? \{link_start}Sign in\{link_end}	¿Ya usas Skype? \{link_start}Inicia sesión\{link_end}	¿Ya está usando Skype? #link_start}Iniciar sesión#link_end}	¿Ya usas Skype? \\link_start}Iniciar sesión\\link_end}	¿Ya usas Skype? \\link_start}Iniciar sesión{link_end}
Close \{topic}'s profile.	Cierre el perfil: \{topic}.	Cerrar el perfil del tema.	Cierra el perfil de \\topic}.	Cerrar el perfil de \\topic}.

Table 7. A selection of the sentences with tags or placeholders translated by three different English → Spanish systems.

The fact that the errors from Table 7 are consistent across the en-es-3 engine suggests that if this engine is used in a professional translation scenario, the error could be solved by a batch search-and-replace operation and the engine could still be useful.

4.1.3. English → Catalan MT Engines

The English → Catalan baseline engine was fine-tuned with a different type of corpus than the others, as described in Section 3.1. Similarly to the English → Spanish engines, the large localization corpus leads to considerable improvement in all four metrics while the small corpus brought a considerable improvement in BLEU (from 38 to 42.5) and TER (from 57.9 to 49.0) and did not lead to any considerable change in chfF2 (63.2 and 63.3 respectively) and COMET (84.3 to 84.6), as may be seen in Figure 4.

system	*comet (μ ± 95% CI)*	*BLEU (μ ± 95% CI)*	*chrF2 (μ ± 95% CI)*	*TER (μ ± 95% CI)*
en-ca-2	84.6 (84.6 ± 2.1)	42.5 (42.3 ± 5.3)	63.3 (63.3 ± 3.7)	49.0 (48.9 ± 6.5)
en-ca-3	**86.6 (86.6 ± 2.0) (p = 0.0030)***	**47.2 (47.0 ± 5.1) (p = 0.0170)***	**67.8 (67.8 ± 3.3) (p = 0.0010)***	**44.0 (43.9 ± 5.7) (p = 0.0230)***

Table 8. A comparison of the en-ca-2 to en-ca-3 MT systems in terms of automatic evaluation metrics. This comparison shows the impact of increasing the fine-tuning corpus from approx. 500,000 words to approx. 2,000,000 in this language pair.

The change from the en-ca-2 engine to en-ca-3 also offers significant improvement across all four metrics. This leads to the conclusion that, similar to the previous engines, English → Catalan performance improves with the addition of further localization data.

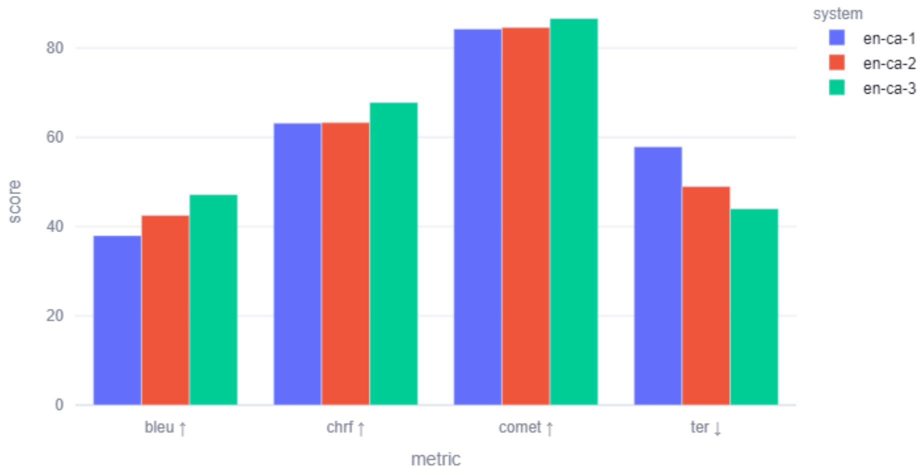


Figure 4. Comparison of the automatic evaluation scores in English → Catalan.

Finally, we overview how English → Catalan engines perform in terms of tags and placeholders and filter the 32 sentences that include tags or placeholders. These segments' average COMET scores are as follows: 81.3, 83.4 and 85.6, as may be seen in Table 8.

source	reference	en-ca-1	en-ca-2	en-ca-3
Avatar of \{title}	Avatar de \{title}	Avatar de \ {títol}	Avatar de \{title}	Avatar de \{title}
Back to \{section}	Torna a \{section}	Torna a \ {secció}	Torna a \{secció}	Torna a \{secció}
There are \{count} participants in conversation	Hi ha \{count} participants a la conversa	Hi ha \ {compte} participants en la conversa	Hi ha \{count} participants en la conversa	Hi ha \{count} participants en la conversa
\{0}Click here\{1} to switch accounts	\{0}Fes clic aquí\{1} per canviar de compte	\ {0} Feu clic aquí\ {1} per canviar de compte	· ·	\{0}Clic aquí \{1} per a commutar comptes
\{ext} File	Fitxer \{ext}	\ {ext} Fitxer	\{ext} Fitxer	\{ext} Fitxer

Table 9. A selection of the sentences with tags or placeholders translated by three different English → Catalan systems.

The en-ca-1 engine tends to translate placeholders and leave space between "\" and the tags. The en-ca-2 engine is inconsistent in terms of translating or retaining the placeholders or tags and, finally, the en-ca-3 engine outputs the tags and placeholders correctly but sometimes continues to translate the placeholders as in the second example in Table 9 ("secció").

4.2. Human Evaluation

For each language pair, the test set of 210 segments translated by each of the three engines was evaluated by three reviewers using the KantanLQR interface. In total, each reviewer rated 630 output sentences. The ranking task used a scale of 3 while the adequacy and fluency tasks used a scale of 5. It was possible to calculate a percentage score for each metric by comparing the total score obtained by an engine against the total possible score. For instance, in the case of ranking, the total possible score is 1890 (reviewer count: 3 × highest score: 3 × segment count: 210). These are represented in percentages in Table 10.

MT Name	Ranking↑	Adequacy↑	Fluency↑
en-tr-1	67.57%	72.35%	75.94%
en-tr-2	77.94%	81.17%	83.37%
en-tr-3	**80.63%**	**82.73%**	**84.22%**
en-es-1	75.03%	80.92%	81.97%
en-es-2	77.25%	83.49%	83.02%
en-es-3	**80.21%**	**84.22%**	**84.38%**
en-ca-1	67.72%	76.35%	73.90%
en-ca-2	72.38%	76.38%	76.79%
en-ca-3	**79.05%**	**80.51%**	**80.25%**

Table 10. Human evaluation results in three metrics displayed in percentages. Percentages in bold are the best scores.

4.2.1. English → Turkish engines

When the three Turkish reviewers evaluated the output from the three engines, in the ranking task the en-tr-1 engine attained 67.57% (1277/1890[14]); en-tr-2 attained 77.94% (1473/1890); and en-tr-3 attained 80.63% (1524/1890) of the overall score. Average ranking scores for each engine are as follows: 2.03, 2.34 and 2.42, respectively. According to these scores, three reviewers rated en-tr-3 as the best performing engine while en-tr-1 obtained the lowest average score.

The adequacy ratings for each engine are as follows: en-tr-1: 72.35% (2279/3150[15]), en-tr-2: 81.17% (2557/3150) and en-tr-3: 82.73% (2606/3150). Average adequacy scores are 3.62, 4.06, and 4.14 respectively. In this evaluation, en-tr-3 obtained the highest score while en-tr-1 obtained the lowest score.

The fluency scores follow a similar pattern as well: en-tr-1: 75.94% (2392/3150), en-tr-2: 83.37% (2626/3150) and en-tr-3: 84.22% (2653/3150). Average fluency scores are 3.80, 4.17 and 4.21. As it can be observed from these scores, en-tr-3 obtained the highest score again while en-tr-1 remained in the lowest position with its score.

Table 11 shows the score averages for each evaluation task and enables investigating the impact of fine-tuning by data augmentation. These average scores suggest that when fine-tuning is conducted by the addition of a bilingual localization corpus, the overall quality improves considerably in this language pair. However, while improvement from en-tr-1 to en-tr-2 or en-tr-1 to en-tr-3 appears dramatic, the quality increase from en-tr-2 to en-tr-3 seems to be minimal. The en-tr-3 engine is fine-tuned with a corpus bigger than the one with en-tr-2; yet this augmentation does not result in an engine with a higher quality. In the specific case of this language pair and domain, we can infer two implications. Firstly, constant data augmentation does not necessarily increase quality in consistently and there may be a plateau after a certain amount of fine-tuning data. Secondly, even a parallel corpus as small as 500,000 source words can provide enough quality improvement to justify the use of the desktop OPUS-CAT MT application with fine-tuning. In the KantanLQR evaluation framework, an adequacy score of 4.06 and a fluency score of 4.17 (from of a scale of 5) may be considered good enough to justify the use of MT for a particular use case.

Av. Scores for En → Tr	*Baseline*	*Small Corpus*	*Large Corpus*
Ranking (max. 3)	2.03	2.34	**2.42**
Adequacy (max. 5)	3.62	4.06	**4.14**
Fluency (max. 5)	3.80	4.17	**4.21**

Table 11. English → Turkish human evaluation results from three reviewers. Best score per metric shown in bold.

4.2.2. English → Spanish engines

The three Spanish evaluators preferred the en-es-3 engine in all the three evaluation tasks. In the ranking task, en-es-1 attained 75.03% (1418/1890), en-es-2 77.25% (1460/1890) and en-es-3 80.21% (1516/1890) of the overall score. The average score for each engine is as follows: 2.25, 2.32 and 2.41. The en-es-3 ranks as the best engine while en-es-1 ranks as the worst.

The adequacy percentages of each engine are 80.92% (2549/3150), 83.49% (2630/3150) and 84.22% (2653/3150). The average scores obtained by each engine reflects these percentages: 4.05, 4.17 and 4.21. The en-es-3 engine has the highest adequacy score while en-es-1 has the lowest one.

Finally, the fluency scores of each engine are 81.97% (2582/3150), 83.02% (2615/3150) and 84.38% (2658/3150). The average fluency scores are 4.10, 4.15 and 4.22 respectively. The en-es-3 obtained the highest score while en-es-1 obtained the lowest score.

Av. Scores for En → Es	*Baseline*	*Small Corpus*	*Large Corpus*
Ranking (max. 3)	2.25	2.32	**2.41**
Adequacy (max. 5)	4.05	4.17	**4.21**
Fluency (max. 5)	4.10	4.15	**4.22**

Table 12. English → Spanish human evaluation results from three reviewers. Best score per metric shown in bold.

Table 12 summarizes the average scores for this language pair. The en-es-3 engine obtained the highest score in all metrics. The human evaluation scores seem to improve gradually from en-es-1 to en-es-3 through data augmentation, as also evidenced by automatic metrics in Figure 3. The improvement from en-es-1 to en-es-2 in all metrics seems to be modest when compared to en-tr engines. Nonetheless, the evolution of the improvements suggests that there is still room for further improvement through the addition of additional corpora. Moreover, the average adequacy and fluency scores of the baseline engine are above four, which indicates that this engine can already provide reasonably good quality results.

4.2.3. English → Catalan

The English → Catalan engines fine-tuned with localization corpora from Softcatalà showed a similar pattern. Of the overall ranking score, the en-ca-1 engine obtained 67.72% (1280/1890), en-ca-2 72.38% (1368/1890), and en-ca-3 79.05% (1368/1890). In parallel to this, average scores for each engine are 2.03, 2.17 and 2.37 respectively. These scores rank en-ca-3 as the best engine while en-ca-1 is the worst.

The adequacy percentage of each engine are 76.35% (2405/3150), 76.38% (2406/3150) and 80.51% (2536/3150). Average adequacy scores are 3.82, 3.82 and 4.03 respectively. With these scores, en-ca-3 is the best performing engine while en-ca-1 and en-ca-2 has the same adequacy scores (with a minimal difference).

Lastly, the fluency percentage of each engine are 73.90% (2328/3150), 76.79% (2419/3150), and 80.25% (2528/3150). Average fluency scores were 3.70, 3.84 and 4.01. According to these results, en-ca-3 is the most fluent engine while en-ca-1 is the least fluent. The average scores may be seen in Table 13.

Av. Scores for En → Ca	*Baseline*	*Small Corpus*	*Large Corpus*
Ranking (max. 3)	2.03	2.17	**2.37**
Adequacy (max. 5)	3.82	3.82	**4.03**
Fluency (max. 5)	3.70	3.84	**4.01**

Table 13. English → Catalan human evaluation results from three reviewers. Best score per metric shown in bold.

Considering the overall results for this language pair, en-ca-3 has the highest scores in all three metrics while en-ca-1 has the lowest except for the adequacy metric. In this metric, en-ca-1 and en-ca-2 share the same score, which indicate that the addition of 500,000 source words did not help improve adequacy. However, data augmentation with 2,000,000M+ source words seems to improve quality since the fluency and adequacy scores passed above four after fine-tuning with the larger corpus.

4.3. Agreement between Reviewers

In this subsection, we firstly focus on the evaluation results per reviewer. Then we share the agreement percentages per engine and aggregated inter-annotator agreement scores per language pair based on Fleiss' Kappa (Fleiss 1971).

For individual reviewer ratings, we report both the percentage scores and the average scores per evaluation metrics. The tables for each language pair are presented in Annex II. Highest scores are highlighted in bold.

For the English→Turkish language pair, the scores given by all three reviewers are compatible with the overall average scores. All of them gave the highest scores to en-tr-3 and the lowest one to en-tr-1 in all three metrics for each engine (see Annex II).

In the Spanish→English language pair, fluency scores agree with the overall results insofar as all three reviewers rank en-es-3 as the most fluent engine. However, in the case of ranking and adequacy, Reviewer 2 gave the same scores for en-es-2 and en-es-3 while the others ranked en-es-3 as the best performing engine.

In the Catalan → English language pair, all three reviewers gave the highest ranking, adequacy and fluency scores to en-ca-3. This implies that there is an overall agreement between the reviewers on the performance of the three engines.

After the individual ratings by the reviewers, we focus on the agreement rates between the reviewers to check the consistency among them. Table 14 shows the percentage of agreement in the ratings in the three evaluation tasks across three MT engines per language pair.

	en-tr-1	**en-tr-2**	**en-tr3**
Ranking	40.95%	39.05%	46.19%
Adequacy	28.10%	33.81%	40.48%
Fluency	18.57%	27.62%	30.48%
	en-es-1	**en-es-2**	**en-es-3**
Ranking	34.76%	34.76%	31.43%
Adequacy	17.62%	17.62%	19.52%
Fluency	28.10%	30.95%	23.81%
	en-ca-1	**en-ca-2**	**en-ca-3**
Ranking	47.14%	44.76%	46.67%
Adequacy	10.95%	14.29%	20.00%
Fluency	9.05%	10.48%	16.19%

Table 14. Percentage of agreement per MT engine in the three human evaluation tasks.

Table 15 provides the aggregated Fleiss' Kappa scores for each language pair and evaluation type. In the following paragraphs, we share the key findings from these two tables per language pair.

	Type of Evaluation	Fleiss' κ
English → Turkish	*Ranking*	0.183
	Adequacy	0.183
	Fluency	0.142
English → Spanish	*Ranking*	0.166
	Adequacy	0.164
	Fluency	0.195
English → Catalan	*Ranking*	0.276
	Adequacy	0.211
	Fluency	0.155

Table 15. Fleiss' Kappa scores per human evaluation type and language pair.

English→Turkish: The inter-annotator agreement scores vary across different MT engines and evaluation types. The agreement scores are highest for Ranking (40.95%, 39.05%, 46.19%), followed by adequacy (28.10%, 33.81%, 40.48%), and lowest for fluency (18.57%, 27.62%, 30.48%). The aggregated Fleiss' Kappa scores are relatively low (0.183 for ranking and adequacy, 0.142 for fluency), suggesting only slight agreement.

English→Spanish: The agreement scores for ranking are fairly consistent across the three MT engines (34.76%, 34.76%, 31.43%). However, the scores for adequacy (17.62%, 17.62%, 19.52%) and fluency (28.10%, 30.95%, 23.81%) show some variation. The Fleiss' Kappa scores are also low (0.166 for ranking, 0.164 for adequacy, 0.195 for fluency), indicating slight agreement.

English→Catalan: This pair has the highest agreement scores for ranking (47.14%, 44.76%, 46.67%) when compared to the other two language pairs. However, the agreement scores for adequacy (10.95%, 14.29%, 20.00%) and fluency (9.05%, 10.48%, 16.19%) are much lower. The Fleiss' Kappa scores reflect a moderate agreement for ranking (0.276) but lower for adequacy (0.211) and fluency (0.155).

5. Discussion and Limitations

The automatic and human evaluations seem to be compatible in terms of the best engines in each language pair when pre-trained NMT models are fine-tuned with localization corpora. Both types of evaluation also suggest that data augmentation has led to a logarithmic-like improvement in English→Turkish output. However, in the case of English→Spanish and English→Catalan, the improvements from each addition were incremental.

The fact that there were improvements suggests that there is still room for more data augmentation in all language pairs and fits with the report by Schwartz *et al.* (2020) that as data sizes increase, added tranches of data become less effective. In the case of English→Catalan, the impact of adding a large corpus has been bigger than in the case of English→Spanish.

If we assume that a score of over 4 in fluency and adequacy ratings justify the use of fine-tuning in OPUS-CAT for professional translators (according to the definitions for each score in KantanLQR; see Table 2), we can make the following arguments based on our localization domain: i) In an English→Turkish localization project, a fine-tuned engine with a small corpus of 500,000 source words may already provide mostly adequate and quite fluent translation results. A further augmentation of the fine-tuning data to 2,000,000+ will have a slight improvement in translation quality; ii) In an English→Spanish localization project, the baseline, a pretrained NMT engine from OPUS-MT, already provides scores over 4 in adequacy and fluency; yet, the quality can be further increased incrementally with the inclusion of 500,000 or 2,000,000 source words of fine-tuning corpora; and iii) In an English → Catalan localization project, fine-tuning with 500,000 source words will not be enough to reach a score of 4 in adequacy and fluency. When fine-tuning is performed with 2,000,000M+ source words, the quality surpasses the score of 4 very slightly but this result hints that a further quality improvement can be achieved through the addition of more fine-tuning data. However, the results in this language pair may be influenced by the quality of the fine-tuning data that was used since the data comes from multiple sources and is therefore expected to be of a more diverse nature. One limitation of the fine-tuning carried out in English→Catalan is that the fine-tuning corpus consisted of localization strings from different open-source projects and the evaluation test data was from a Microsoft project unlike other language pairs which were fine-tuned and evaluated with Microsoft corpora.

Aside from the aforementioned limitation, our study has other limitations regarding fine-tuning in OPUS-CAT as well as evaluation resulting from methodological preferences. Nieminen (2021: 214) states that it is possible to continue fine-tuning for multiple epochs and modify learning rates. However, these modifications can increase fine-tuning durations and may also lead to overfitting as observed by the author. Long durations of fine-tuning may not be optimum for professional translators who need to create an engine rapidly for a translation project with a tight deadline. Overfitting leads to engines that memorise the training data at the expense of losing generalisation capabilities, resulting in low quality translations. Further studies can be made using the same corpus and changing fine-tuning settings in each iteration. When it comes to evaluation limitations, localization strings are usually short and context-dependent; one string may have multiple translations depending on the context. We utilized one single test file to maximise overall consistency, but we removed in-file and cross-file repetitions, which constrained the context of the file. In our evaluation design, reviewers only saw the source string with its three possible translations within KantanLQR. We increased the number of test strings to compensate for this limitation.

6. Conclusion

Our study showed that it is possible to achieve significant translation quality improvements over pretrained NMT models in three language pairs fine-tuned with specific domain corpora. These results were achieved in a desktop Windows environment without the need to connect to an external server. While it should be noted that fine-tuning does not necessarily guarantee this outcome in every corpus size and type, it can be argued that when confidentiality and privacy are of high concern, fine-tuned, desktop-based engines can be a viable alternative to commercial systems for professional translators. Furthermore, the use of pretrained NMT engines removes the need for costly MT training and provides a more environmentally friendly alternative since less energy is consumed.

Applications such as OPUS-CAT and pretrained models coming from OPUS-MT project lay the foundation for a future where the translator is not only a passive user of MT systems but also an empowered professional who is able to make informed decisions about how and when to use MT in their workflow. This removes an element of control from the client or translation employer, but also removes the client-side cost of MT training and preparation.

Availability of free and open desktop MT applications as well as pretrained models can potentially empower translators. Moreover, when and if a translator can combine these technologies with high quality, specific domain TMs, productivity gains can be increased. Hence, having free and open domain-specific parallel corpora in very different language pairs is essential. Extending the capabilities of OPUS-CAT (and other similar future applications) to include operating systems other than Windows will be important as well. Finally, adding more features to OPUS-CAT and other similar toolkits can lead to other productive ways of using desktop MT.

In the future, we would like to compare our best engines with commercial systems such as Google Translate to study their relative quality. Another line of study could be to train professional translators to use OPUS-CAT and collect data about their perceptions about the usability of the application within their professional workflows.

Acknowledgements

This work is supported by the European Union-NextGenerationEU grant in the framework of Margarita Salas Postdoctoral Grant and also with the financial support of Science Foundation Ireland at ADAPT, the SFI Research Centre for AI-Driven Digital Content Technology at Dublin City University [13/RC/2106_P2]. We would like to thank the reviewers who kindly agreed to participate in the evaluation: Muhammed Baydere, Emre Canbaz, Şevval Mutlu, Eden Y. Bravo Montenegro, Mer Moreno, Ariana López Pereira, Xènia Amorós, Eduard Simón and Marc Riera. We are also grateful to KantanAI for providing us access to their platform for research purposes.

References

- **Abdallah, Kristiina** (2012). *Translators in production networks: Reflections on agency, quality and ethics*. University of Eastern Finland.

- **Cadwell, Patrick, Sharon O'Brien and Carlos SC Teixeira** (2018). "Resistance and accommodation: Factors for the (non-) adoption of machine translation among professional translators." *Perspectives*, *26*(3), 301–321. https://doi.org/10.1080/0907676X.2017.1337210

- **Esselink, Bert** (2003). "Localisation and translation" Harold Somers (ed.), *Computers and Translation: A translator's guide.* Amsterdam: John Benjamins, 67–86. https://doi.org/10.1075/btl.35.08ess

- **Fleiss, Joseph L.** (1971). "Measuring nominal scale agreement among many raters." *Psychological Bulletin*, 76(5), 378–382.

- **Herbig, Nico, Tim Düwel, Santanu Pal, Kalliopi Meladaki, Mahsa Monshizadeh, Antonio Krüger and Josef van Genabith** (2020). "MMPE: A Multi-Modal Interface for Post-Editing Machine Translation." *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1691–1702. https://doi.org/10.18653/v1/2020.acl-main.155

- **Jiménez-Crespo, Miguel A.** (2020). "The "technological turn" in translation studies: Are we there yet? A transversal cross-disciplinary approach." *Translation Spaces*, *9*(2), 314–341. https://doi.org/10.1075/ts.19012.jim

- **Kenny, Dorothy and Stephen Doherty** (2014). "Statistical machine translation in the translation curriculum: Overcoming obstacles and empowering translators." *The Interpreter and Translator Trainer*, *8*(2), 276–294. https://doi.org/10.1080/1750399X.2014.936112

- **Kenny, Dorothy** (ed.) (2022). *Machine translation for everyone: Empowering users in the age of artificial intelligence*. Berlin: Language Science Press. https://doi.org/10.5281/ZENODO.6653406

- **Martín Mor, Adrià** (2017). "MTradumàtica: Statistical machine translation customisation for translators." *Skase. Journal of Translation and Interpretation*, *11*(1), 25–40.

- **Moorkens, Joss** (2022). "Ethics and Machine Translation". Dorothy Kenny (ed.) (2022). *Machine translation for everyone.* Berlin: Language Science Press, 121-140. https://doi.org/10.5281/zenodo.6759984

- **Moorkens, Joss and David Lewis** (2019). "Copyright and the reuse of translation as data." Minako O'Hagan (ed) (2019). *The Routledge Handbook of Translation and Technology.* Abingdon: Routledge, 469–481. http://dx.doi.org/10.4324/9781315311258-28

- **Moorkens, Joss and Sharon O'Brien.** (2017). "Assessing User Interface Needs of Post-Editors of Machine Translation". Dorothy Kenny (ed.) (2017). *Human Issues in Translation Technology: The IATIS Yearbook*. Abingdon: Routledge, 109-130.

- **Nieminen, Tommi** (2021). "OPUS-CAT: Desktop NMT with CAT integration and local fine-tuning." *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 288–294. https://doi.org/10.18653/v1/2021.eacl-demos.34

- **Papineni, Kishore, Salim Roukos, Todd Ward and Wei-Jing Zhu** (2002). "BLEU: A method for automatic evaluation of machine translation." Pierre Isabelle *et al.* (eds) (2002). *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics.* Philadelphia: Association for Computational Linguistics, 311-318.

- **Popović, Maja** (2015). "chrF: Character n-gram F-score for automatic MT evaluation." *Proceedings of the Tenth Workshop on Statistical Machine Translation*, 392–395. https://doi.org/10.18653/v1/W15-3049

- **Rei, Ricardo, Craig Stewart, Ana C. Farinha and Alon Lavie** (2020). "COMET: A Neural Framework for MT Evaluation." *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2685–2702. https://doi.org/10.18653/v1/2020.emnlp-main.213

- **Ragni, Valentina and Lucas Nunes Vieira** (2022). "What has changed with neural machine translation? A critical review of human factors." *Perspectives*, *30*(1), 137–158. https://doi.org/10.1080/0907676X.2021.1889005

- **Ramírez-Sánchez, Gema, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Caroline Rossi, Dorothy Kenny, Riccardo Superbo, Pilar Sánchez-Gijón and Olga Torres-Hostench** (2021). "MultiTraiNMT: Training Materials to Approach Neural Machine Translation from Scratch." *Proceedings of the Translation and Interpreting Technology Online Conference*, 181–186. https://aclanthology.org/2021.triton-1.21

- **Ramos, María del Mar Sánchez, Jesús Torres-del-Rey and Lucía Morado.** (2022). "Localisation Training in Spain and Beyond: Towards a Consensus on Content and Approach." *Hermes*, 62, 1-26. https://doi.org/10.7146/hjlcb.vi62.128626

- **Schwartz, Roy, Jesse Dodge, Noah A. Smith and Oren Etzioni** (2020). "Green AI." *Communications of the ACM*, *63*(12), 54–63. https://doi.org/10.1145/3381831

- **Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla and John Makhoul** (2006). "A Study of Translation Edit Rate with Targeted Human Annotation." *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, 223–231. https://aclanthology.org/2006.amta-papers.25

- **Tiedemann, Jörg** (2012). "Parallel Data, Tools and Interfaces in OPUS." Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk and Stelios Piperidis (eds) (2012) *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12),* 2214–2218. Istanbul: LREC.

- **Tiedemann, Jörg, Mikko Aulamo, Daria Bakshandaeva, Michele Boggia, Stig-Arne Grönroos, Tommi Nieminen, Alessandro Raganato, Yves Scherrer, Raul Vazquez and Sami Virpioja** (2022). "Democratizing Machine Translation with OPUS-MT." arXiv:2212.01936. http://arxiv.org/abs/2212.01936

- **Tiedemann, Jörg and Santhosh Thottingal** (2020). "OPUS-MT – Building open translation services for the World." *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, 479–480. Lisboa: EAMT. https://aclanthology.org/2020.eamt-1.61

- **Vanroy, Bram, Arda Tezcan and Lieve Macken** (2023). "MATEO: MAchine Translation Evaluation Online." *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, 499–500. Tampere, Finland: EAMT.

Annex I. Task Instructions for Reviewers

Guidelines for performing the Machine Translation Evaluation Task in the Localization Domain

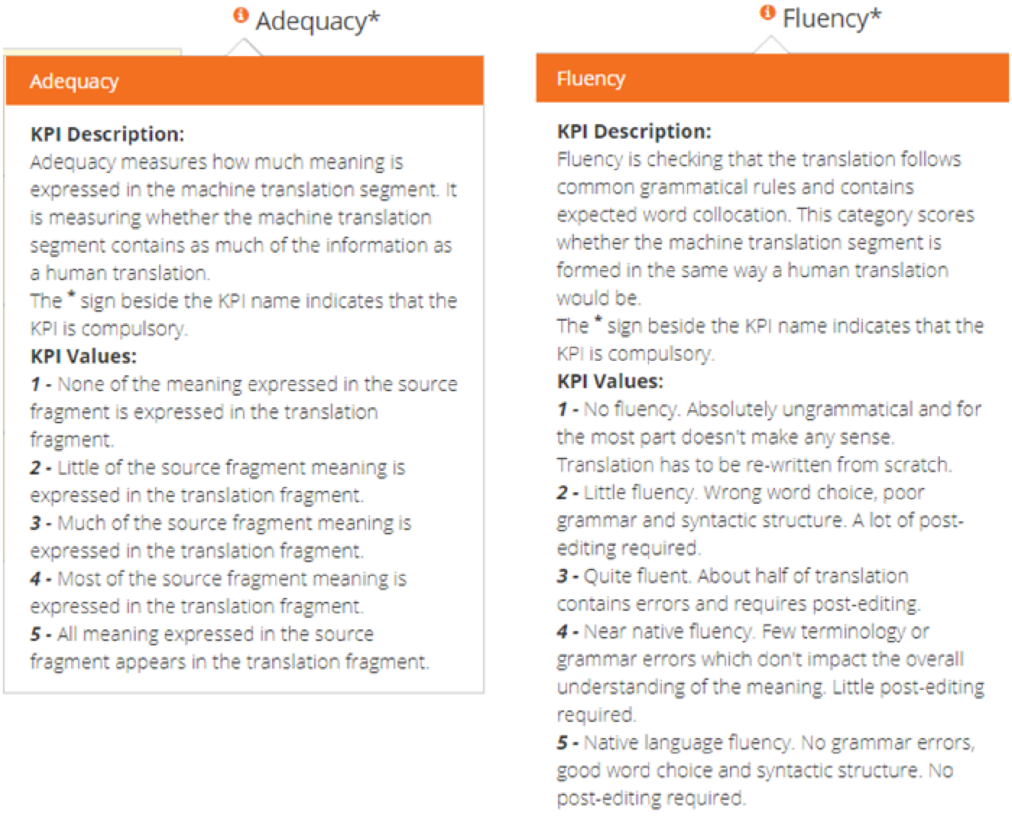
Title:	Finetuning Machine Translation Engines with Custom Parallel Corpus and Possible Quality Improvements
Objective:	The task consists of evaluating the translation quality of 3 Machine Translation Engines in 3 language pairs by human reviewers. The ultimate objective of the study is to measure if a desktop MT application (OpusCAT) finetuned with different sizes of custom localization corpus can provide significant translation quality to make it useful for translators.
Task facilitator:	Gokhan Dogru (Postdoctoral Researcher, UAB & DCU)
Organization:	Facultat de Traducció i d'Interpretació, Universitat Autònoma de Barcelona - School of Applied Language & Intercultural Studies, Dublin City University
Project By:	Gokhan Dogru & Joss Moorkens
Approval	This experiment is approved by the Ethics Commission on Animal and Human Experimentation with the number of 20190927CEEEAH.
Dates	07.11.2022 - 14.11.2022

Task Guidelines

1. Please fill out the short survey aiming at collecting professional details of the participants. It should take less than 3 minutes to complete. Form link is here: https://forms.gle/kiA51ehucXzcPvtb9
2. Once you complete the survey, we will send you a link to your email address to connect to KantanLQR platform. You will need to enter with your email and create a password (if you haven't done so before).
3. Once you login to the platform, you will see the dashboard with the task. You should first click on the "?" to accept the task. Once you accept the task, you can click on the pen symbol and begin the evaluation task. The strings are from the interface of Skype web application. In case anything is not clear, you can use the comment section to write your comment. But it is not obligatory.

4. In the upper left corner, you will see the source sentence, and below it there will be its 3 machine translations by different MT engines. The order of these translations is randomized in each step to avoid bias.
5. There are 2 evaluation criteria: adequacy and fluency. In a nutshell, adequacy measures the accuracy of the translation compared to the source sentence while fluency measures how grammatically correct the translation. You will have a scale of 5 stars. More stars mean better adequacy or fluency. The "i" symbol near each title gives hints about the meaning of each star and the respective definitions. See the image (i) below.
6. Finally, you are expected to rank each engine from the best to the worst. Again, more stars mean better quality. Hence, the best translation result should get 3 stars while the worst one should get 1 star. Note that if you think two engines are equal, you can assign the same number of stars to them.
7. You can press "Finish" to pause and leave the task before finishing and return later to complete it. There are 210 mostly very short source sentences to be evaluated.
8. If you need more help about the task, please send an email to gokhan.dogru@uab.cat.
9. A simulation of the steps from the reviewer's perspective are also available in a video: https://drive.google.com/file/d/1bNbTVUhdvJDvenVryHuHIyFjoUFkgBh6/view?usp=sharing

(i) **Image**: Tips for understanding Adequacy and Fluency available in KantanLQR



Adequacy

KPI Description:
Adequacy measures how much meaning is expressed in the machine translation segment. It is measuring whether the machine translation segment contains as much of the information as a human translation.
The * sign beside the KPI name indicates that the KPI is compulsory.
KPI Values:
1 **-** None of the meaning expressed in the source fragment is expressed in the translation fragment.
2 **-** Little of the source fragment meaning is expressed in the translation fragment.
3 **-** Much of the source fragment meaning is expressed in the translation fragment.
4 **-** Most of the source fragment meaning is expressed in the translation fragment.
5 **-** All meaning expressed in the source fragment appears in the translation fragment.

Fluency

KPI Description:
Fluency is checking that the translation follows common grammatical rules and contains expected word collocation. This category scores whether the machine translation segment is formed in the same way a human translation would be.
The * sign beside the KPI name indicates that the KPI is compulsory.
KPI Values:
1 **-** No fluency. Absolutely ungrammatical and for the most part doesn't make any sense. Translation has to be re-written from scratch.
2 **-** Little fluency. Wrong word choice, poor grammar and syntactic structure. A lot of post-editing required.
3 **-** Quite fluent. About half of translation contains errors and requires post-editing.
4 **-** Near native fluency. Few terminology or grammar errors which don't impact the overall understanding of the meaning. Little post-editing required.
5 **-** Native language fluency. No grammar errors, good word choice and syntactic structure. No post-editing required.

Annex II

Human evaluation results according to reviewer:

English → Turkish

EN → TR	***Ranking***			***Adequacy***			***Fluency***		
Percentage	**Baseline**	**Small Corpus**	**Big Corpus**	**Baseline**	**Small Corpus**	**Big Corpus**	**Baseline**	**Small Corpus**	**Big Corpus**
reviewer 1	73.81%	84.76%	**87.94**%	80.76%	88.19%	**89.90**%	87.33%	92.48%	**92.57**%
reviewer 2	64.13%	73.81%	**75.56**%	60.29%	72.10%	**73.05**%	60.57%	71.24%	**72.86**%
reviewer 3	64.76%	75.24%	**78.41**%	76.00%	83.24%	**85.24**%	79.90%	86.38%	**87.24**%
Average	**Baseline**	**Small Corpus**	**Big Corpus**	**Baseline**	**Small Corpus**	**Big Corpus**	**Baseline**	**Small Corpus**	**Big Corpus**
reviewer 1	2.22	2.54	**2.65**	4.05	4.41	**4.51**	4.37	4.62	**4.64**
reviewer 2	1.93	2.22	**2.27**	3.02	3.60	**3.67**	3.04	3.57	**3.66**
reviewer 3	1.95	2.25	**2.36**	3.80	4.16	**4.27**	3.99	4.32	**4.36**

English → Spanish

EN → ES	***Ranking***			***Adequacy***			***Fluency***		
Percentage	**Baseline**	**Small Corpus**	**Big Corpus**	**Baseline**	**Small Corpus**	**Big Corpus**	**Baseline**	**Small Corpus**	**Big Corpus**
reviewer 1	80.48%	83.33%	**89.52** %	87.14%	89.43%	**90.57** %	82.86%	84.76%	**87.05** %
reviewer 2	71.90%	**74.76**%	**74.76** %	75.62%	**78.00**%	**78.00** %	78.67%	79.90%	**80.57** %
reviewer 3	72.70%	73.65%	**76.35** %	80.00%	83.05%	**84.10** %	84.38%	84.38%	**85.52** %
Average	**Baseline**	**Small Corpus**	**Big Corpus**	**Baseline**	**Small Corpus**	**Big Corpus**	**Baseline**	**Small Corpus**	**Big Corpus**
reviewer 1	2.42	2.50	**2.68**	4.36	4.47	**4.53**	4.14	4.23	**4.35**
reviewer 2	2.16	**2.25**	**2.25**	3.79	**3.91**	**3.91**	3.93	3.99	**4.02**
reviewer 3	2.19	2.22	**2.30**	4.01	4.17	**4.22**	4.23	4.23	**4.29**

English → Catalan

EN → CA	***Ranking***			***Adequacy***			***Fluency***		
Percentage	**Baseline**	**Small Corpus**	**Big Corpus**	**Baseline**	**Small Corpus**	**Big Corpus**	**Baseline**	**Small Corpus**	**Big Corpus**

reviewer 1	67.94%	69.68%	**72.54%**	75.33%	74.00%	**77.43%**	70.29%	71.52%	**73.90%**
reviewer 2	65.08%	69.21%	**79.84%**	69.81%	70.10%	**77.14%**	66.57%	70.86%	**76.76%**
reviewer 3	70.16%	78.25%	**84.76%**	83.90%	85.05%	**86.95%**	84.86%	88.00%	**90.10%**
Average	**Baseline**	**Small Corpus**	**Big Corpus**	**Baseline**	**Small Corpus**	**Big Corpus**	**Baseline**	**Small Corpus**	**Big Corpus**
reviewer 1	2.04	2.09	**2.18**	3.76	3.69	**3.87**	3.51	3.57	**3.69**
reviewer 2	1.96	2.08	**2.40**	3.50	3.52	**3.87**	3.34	3.56	**3.85**
reviewer 3	2.10	2.35	**2.55**	4.19	4.25	**4.34**	4.24	4.40	**4.51**

Biographies

Gokhan Dogru is a visiting postdoctoral researcher at ADAPT-DCU affiliated with the Faculty of Translation and Interpreting at Universitat Autònoma de Barcelona (UAB) in the framework of a Margarita Salas Grant. His research interests include terminological quality evaluation in machine translation, different use cases of MT for professional translators and the intersection of translation profession and translation technologies as well as localization.

ORCID ID: 0000-0001-7141-2350

E-mail: gokhan.dogru@uab.cat

Joss Moorkens is an Associate Professor at the School of Applied Language and Intercultural Studies in Dublin City University (DCU), Challenge Lead at the ADAPT Centre, and member of DCU's Institute of Ethics, and Centre for Translation and Textual Studies. He is General Co-Editor of Translation Spaces with Dorothy Kenny and coauthor of the textbooks *Translation Tools and Technologies* (Routledge 2023) *Translation Automation* (Routledge 2024).

ORCID ID: 0000-0003-0766-0071

E-mail: joss.moorkens@dcu.ie

Notes

[1] https://omegat.org/ (last access: 03.11.2023)
[2] https://www.trados.com/ (last access: 03.11.2023)
[3] https://www.memoq.com/ (last access: 03.11.2023)
[4] Microsoft Visual Studio's *Translation and UI Strings*: https://my.visualstudio.com/downloads?pid=6822 (last access: 07.11.2022)
[5] KantanLQR. https://kantanmt.zendesk.com/hc/en-us/articles/115003644483-What-is-KantanLQR- (last access: 04.09.2023)
[6] https://mtradumatica.uab.cat/ (last access: 03.11.2023)
[7] https://ntradumatica.uab.cat/ (last access: 03.11.2023)
[8] English → Turkish OPUS-MT Model. https://github.com/Helsinki-NLP/Tatoeba-Challenge/tree/master/models/eng-tur (last access: 09.11.2022)
[9] English → Spanish OPUS-MT Model. https://github.com/Helsinki-NLP/Tatoeba-Challenge/tree/master/models/eng-spa (last access: 09.11.2022)
[10] English → Catalan OPUS-MT Model. https://github.com/Helsinki-NLP/Tatoeba-Challenge/tree/master/models/eng-cat (last access: 09.11.2022)
[11] Instead of downloading the largest translation memory, the one including all Softcatalà projects (Totes les memòries de projectes de Softcatalà) is downloaded for translation quality concerns. https://www.softcatala.org/recursos/memories/ (last access: 07.11.2022)
[12] https://github.com/gokhandogru/MT-Fine-tuning-for-Turkish-Spanish-and-Catalan
[13] MATEO. https://mateo.ivdnt.org/ (last access: 14.09.2023)
[14] 210 sentences × a rating over a scale of 3 × 3 reviewers = 1890
[15] 210 sentences × a rating over a scale of 5 × 3 reviewers = 3150