

English-Catalan Neural Machine Translation: state-of-the-art technology, quality, and productivity

Vicent Briva-Iglesias



Vicent Briva-Iglesias
Dublin City University;
vicent.brivaiglesias2@mail.dcu.ie;
ORCID:
[0000-0001-8525-2677](https://orcid.org/0000-0001-8525-2677)



Abstract

Recent major changes and technological advances have consolidated machine translation (MT) as a key player to be considered in the language services world. In numerous instances, it is even an essential player due to budget and time constraints. Much attention has been paid to MT research recently, and MT use by professional or amateur users has increased. Yet, research has focused mainly on language combinations with huge amounts of online available corpora (e.g. English-Spanish). The situation for minoritized or stateless languages like Catalan is different. This study analyses Softcatalà's new open-source, neural machine translation engine and compares it with Google Translate and Apertium in the English-Catalan language pair. Although MT engine developers use automatic metrics for MT engine evaluation, human evaluation remains the gold standard, despite its cost. Using TAUS DQF tools, translation quality (in terms of relative ranking, adequacy and fluency) and productivity (comparing editing times and distances) have been evaluated with the participation of 11 evaluators. Results show that Softcatalà's Translator offers higher quality and productivity than the other engines analysed.

Keywords: machine translation, translation technologies, human evaluation, Catalan, translation quality, quality evaluation.

Resumen

Los recientes e importantes cambios y avances tecnológicos han consolidado la traducción automática (TA) como un actor clave a tener en cuenta en el mundo de los servicios lingüísticos. En numerosos casos, es incluso un actor esencial debido a las limitaciones de presupuesto y tiempo. Últimamente se ha prestado mucha atención a la investigación en TA y ha aumentado su uso por parte de usuarios profesionales y aficionados. Sin embargo, la investigación se ha centrado principalmente en las combinaciones lingüísticas con grandes cantidades de corpus disponibles en línea (por ejemplo, inglés-español). La situación de las

lenguas minoritarias o no oficiales en un estado, como el catalán, es distinta. Este estudio analiza el nuevo motor de traducción automática neuronal de código abierto de Softcatalà y lo compara con Google Traductor y Apertium en la combinación lingüística inglés-catalán. Aunque los desarrolladores de motores de traducción automática utilizan métricas automáticas para su evaluación, la evaluación humana sigue siendo la práctica de referencia, a pesar de su coste. Mediante las herramientas TAUS DQF, se ha evaluado la calidad de la traducción (en términos de clasificación relativa, adecuación y fluidez) y la productividad (comparando los tiempos de edición y las distancias) con la participación de 11 evaluadores. Los resultados muestran que el traductor de Softcatalà ofrece mayor calidad y productividad que los otros motores analizados.

Palabras clave: traducción automática, tecnologías de la traducción, evaluación humana, catalán, calidad de la traducción, evaluación de la calidad.

Resum

Els recents i importants canvis i avenços tecnològics han consolidat la traducció automàtica (TA) com un actor clau a tenir en compte en el món dels serveis lingüístics. En molts casos, és fins i tot un actor essencial a causa de les limitacions de pressupost i temps. Últimament, la recerca en TA ha rebut molta atenció i se n'ha augmentat l'ús per part d'usuaris professionals i aficionats. De tota manera, la recerca s'ha centrat principalment en les combinacions lingüístiques amb grans quantitats de corpus disponibles en línia (per exemple, anglès-castellà). La situació de les llengües minoritàries o no oficials a un estat, com el català, és diferent. Aquest estudi analitza el nou motor de traducció automàtica neuronal de codi obert de Softcatalà i el compara amb el Google Traductor i l'Apertium en la combinació lingüística anglès-català. Tot i que els desenvolupadors de motors de traducció automàtica fan servir mètriques automàtiques per avaluar-los, l'avaluació humana continua sent la pràctica de referència, tot i el cost que implica. Per mitjà de les eines TAUS DQF, s'ha avaluat la qualitat de la traducció (en termes de classificació relativa, adequació i fluïdesa) i la productivitat (comparant els temps d'edició i les distàncies) amb la participació d'11 avaluadors. Els resultats mostren que el traductor de Softcatalà ofereix una qualitat i productivitat majors que els altres motors analitzats.

Paraules clau: traducció automàtica, tecnologies de la traducció, avaluació humana, català, qualitat de la traducció, avaluació de la Qualitat

1. Introduction

Recent major changes and technological advances have consolidated MT as a key player to be considered in the language services industry (whether or not we are talking about translation, localization or internationalization, among other services). In numerous instances, MT has even become an essential player due to budget and time constraints. This has led to continuous, widespread and profound changes in the professional and academic world of translation (Cronin, 2012).

With current technological advances, the use of artificial intelligence, and the development of deep learning, the power of computers and machines has increased substantially, which has allowed language service providers to reduce the processing time of many tasks with automation and to increase translators' productivity (Ahrenberg, 2017). In this context, neural machine translation (NMT) offers better results than previous systems or paradigms (Bentivogli et al., 2016). According to Wu et al. (2016: 2), "the quality of the resulting [neural machine] translation system gets closer to that of average human translators." Yet, Läubli et al. (2018) and Toral (2020) have rejected different

claims stating that NMT offered human-quality translation or even better translation quality than humans, and highlighted that these claims were based on biased evaluations, i.e., these evaluations were either automatic or performed by non-professional translators. These authors also highlighted the utmost importance of performing human evaluations of MT systems, in accordance with Läubli et al. (2020).

Most research in MT normally studies a small group of language pairs (e.g., English, German, Spanish or French), which we can call high-resourced languages, i.e., languages with huge amounts of parallel corpora available on the web. This article analyses open-source tools, but with the aim of evaluating NMT engines for the English-Catalan language pair. To do this, a study was carried out to evaluate the quality and performance of a new open-source NMT engine created by Softcatalà, and compare it with its predecessor, Apertium (a rule-based MT engine) and Google's proprietary NMT engine.

Firstly, the choice of Google Translate is because it is the best known and most widely used commercial and proprietary MT engine to date, and is therefore the one used by most students or MT users (Pitman, 2021). Secondly, Apertium is the engine most used by Softcatalà members in their free/open-source software localization collaborations, and they have shown their interest in finding out whether it is worth changing their MT system from Apertium to Softcatalà's Translator to increase their productivity. Therefore, the aim of this article can be summarized as follows:

1. To analyse which MT method and system of choice, either Apertium (open-source rule-based MT), Softcatalà's Translator (open-source NMT) or Google Translate (proprietary NMT), offered a higher translation quality.
2. To find out which MT engine offered the best post-editing performance.

2. Machine Translation Quality Evaluation

In the language services industry, MT interest rests mainly on two principles: increasing productivity and reducing costs (normally, from the perspective of language service providers). There is currently a lot of material to translate and customers want documents to be translated faster and cheaper (EUATC, 2020). When creating an MT engine, one of the main goals of MT developers in an industry setting is to be able to translate more content in less time, that is, translators can translate faster with MT aids. To discover whether this is possible, MT developers need to corroborate and check if their MT engine is performing correctly, as well as whether the modifications that have been introduced have served to improve the system. Also, translators, language service providers, and MT users need to ensure that translations meet minimum quality criteria. To achieve this, MT quality evaluation is required, and has received a lot of attention in the literature lately (Barrault et al., 2020b, 2021). We can identify two forms of evaluation widely accepted by academia and industry: automatic evaluation and human evaluation.

Automatic MT evaluation is probably the most used form of MT evaluation because of its reduced cost and fast results. According to Martín-Mor et al. (2016: 63–64), automatic MT evaluation:

...focuses on refining quality indices by comparing MT raw translations with human reference translations (also known as a gold standard), or with a comparable corpus of texts in the target language. If a human translation of the same text exists, each segment is compared in terms of the number of edits (insertions, deletions and substitutions) needed to convert each segment of the raw translation into the segment of the human translation. [Translation by the author]

Although many automatic quality assessment metrics have emerged, there are two main groups of automatic MT evaluation metrics. On the one hand, those that measure the edit distance, i.e., how many changes are necessary to transform the raw MT text into the gold standard sentence used as the reference, like WER (Ye-Yi Wang et al., 2003), PER (Tillmann et al., 1997) or TER (Snover et al., 2016, 2009). On the other hand, those that focus on the order of words or groups of words. Within this latter group of metrics, BLEU (Papineni et al., 2001) has been the norm in automatic MT evaluation in recent decades. BLEU compares the n-grams of the raw MT output with the n-grams of the gold standard translation used as a reference and counts the number of matches. As the matches are position-independent, the higher the matches, the better the candidate translation is supposed to be. In other words, these types of metrics calculate the frequency with which words or phrases (up to a set of 4 words in the case of BLEU) match in both the human reference and the proposed translation by the MT engine.

The main problem of automatic MT evaluation is that one sentence can be translated into another language in many different and valid ways. Thus, sentences deviating in style, but retaining the meaning from the source language, are penalized by automatic MT evaluation metrics. Language is complex, and its inner rules make language evaluation a difficult issue to solve automatically. Recent studies have claimed that using only automatic MT evaluation metrics to decide the superiority of one MT system over another is misleading the research field of MT development (Freitag et al., 2021, 2020; Kocmi et al., 2021), and have proposed that COMET (Rei et al., 2020) and chrF (Popović, 2015) may be the best automatic metrics to date in terms of measuring quality. It is worth stressing that, even though automatic MT evaluation metrics are not the best way forward, they are cheap, fast, reproducible and give some valuable feedback, which may then be accompanied by the widely-accepted, best method to evaluate MT: human evaluation.

Läubli et al. (2020: 1–2) pointed out that "human evaluation remains the gold standard, but there are many design decisions that potentially affect the validity of such a human evaluation" and that "there is consensus that a reliable evaluation should (despite high costs) be carried out by humans." Therefore, designing a good methodology to evaluate MT has been one of the main objectives in the translation technology sector in recent decades, as can be seen in the multiple annual conferences dealing with MT quality evaluation methodologies (Barrault et al., 2020a). Multiple methods for human evaluation of MT have been proposed (Moorkens et al., 2018): scoring segments from 1 to 5 for quality conformance; making a relative ranking; analysing errors by MQM-DQF metrics (Görög, 2014); by comprehension testing; or post-editing. However, two methods stand out:

- Relative ranking: evaluators are shown the original sentence in the source language and several choices of MT engines in the target language. Then, they

rank relatively which engines offer better translation proposals. E.g., engine A is the best, engine B is the second best, and engine C is the worst (Bojar et al., 2016; Koehn and Monz, 2006).

- Direct evaluation: evaluators are shown the original sentence in the source language and go through the translated sentences (either different MT engines or human translations) one at a time. Then, they have to assign a score from 0 to 100 to each translation, or assign the error type and the penalty level of the mistake. This method allows evaluators to find out which engine is better and also indicates to what degree a specific engine is better (Briva-Iglesias, 2021). To avoid problems of subjectivity and scoring between the different evaluators, documents with instructions are previously prepared to homogenize the criteria (Graham et al., 2013).

According to Callison-Burch et al. (2007), the evaluation by relative ranking provides a higher agreement among the evaluators than the other methods on a best to worst scale. However, it is not possible to know to what extent certain engines are better since the evaluation is only relative.

3. Methodology

This section defines the methodology followed to achieve the objectives of this study, which has been based on similar work in the English-Spanish language combination (López-Pereira, 2019). This study started from the hypothesis that NMT engines would offer a higher translation quality (in terms of adequacy and fluency) than the rule-based MT engine. However, we did not know which NMT engine (Softcatalà's Translator or Google Translate) would offer higher quality. In addition, rule-based MT may produce clearer errors, but these will always be repeated and may be easier detect and edit. Castilho et al. (2017) indicated that NMT systems achieved fluent and grammatically correct sentences, even though meaning errors may also be present. For this reason, NMT may be a double-edged sword and, when it comes to post-editing, there was a possibility that NMT would require more time to find the underlying errors and achieve the output we were looking for in comparison with rule-based MT.

3.1. MT Systems

This study evaluated three different MT systems. First, Apertium¹, which is a rule-based MT engine that was designed within a public funding research project (Forcada et al., 2011).

Second, Softcatalà's Translator², which is an open-source NMT engine that was released and built by Softcatalà in mid-2020. Softcatalà is a non-profit organization focused on promoting the use of Catalan in computing, Internet and new technologies. This engine was exclusively trained with a corpus of translations of free and open-source products. The English-Catalan model analysed in this paper was the first release of Softcatalà's Translator system, and data were studied in June 2020.

¹ Please see <https://www.apertium.org/>, last accessed on 21st September 2021.

² Please see <https://www.softcatala.org/traductor/>, last accessed on 21st September 2021.

Third, Google Translate, a proprietary NMT engine developed by Google. Google Translate has been using NMT since the early 2020s, and it is the most widely used MT system in the world (Slator, 2016). Yet, Google Translate is proprietary, and therefore translations are retained within the system and there may be confidentiality and privacy issues. The MT raw output analysed in this paper was produced in June 2020.

3.2. Evaluators

According to Läubli et al.'s (2020) research, evaluator selection is a crucial step when defining the methodology. Human evaluation can be carried out by both professional and amateur translators. However, Castilho et al. (2018: 23) indicated that there was a tendency to, “rely on students and amateur evaluators, sometimes with an undefined (or self-rated) proficiency in the languages involved, an unknown expertise with the text type,” as it was easier to get their collaboration. Castilho et al. (2018) also highlighted that non-expert translators lacked knowledge of translation, and therefore may not notice subtle differences that made one translation more suitable than another. Thus, when confronted with a translation that is hard to post-edit, non-expert translators tended to accept the MT rather than try to improve it.

Consequently, careful consideration was given to the selection of evaluators, and one of the guiding objectives was that evaluators should have similar knowledge, otherwise the results obtained could vary greatly, depending on the speciality and experience of each person. In the following section, a more thorough explanation on the profiles of the evaluators for each test is given.

3.3. Human Evaluation

As automatic metrics do not guarantee a complete and appropriate quality assessment of MT systems, performing a human evaluation is the norm in today's industry (Freitag et al., 2021). To perform such an evaluation in this paper, the TAUS Dynamic Quality Framework (DQF) tools³ were used. To avoid evaluator subjectivity or to reduce it as far as possible, evaluation guidelines were designed which had to be read carefully by the evaluators before carrying out the different tests. These guidelines served to homogenise the assessment criteria of the evaluators and can be found in Annex 1 (Guidelines for the MT Ranking evaluation), Annex 2 (Guidelines for the Adequacy and Fluency evaluations), and Annex 3 (Guidelines for the Productivity evaluation). The different human evaluation tests were carried out in the following order.

MT Ranking

In this first test, a relative ranking evaluation was performed. Evaluators were presented with a sentence in the source language together with three translation proposals. These proposals corresponded to each of the MT systems (Apertium, Softcatalà's Translator and Google Translate) and were anonymous, so evaluators did not know which MT system originated each of the proposals. Then, evaluators had to classify which

³Please see <https://dqf.taus.net/>, last accessed on 21st September 2021.

translations were the best by assigning a score from 1 to 3. For example: option A is the best (1), option B is the second best (2), and option C is the worst (3).

For the MT Ranking test, 11 professional translators participated in the evaluation. All of them had a degree in Translation and Interpreting, a Master's degree in Specialised Translation, and had Catalan as their mother tongue. In addition, all of them had 1 to 3 years of professional experience in the language services industry. This profile was chosen because these participants had translation training, sufficient technical knowledge to be able to evaluate a specialised translation text in the software localisation domain, and professional experience.

Fluency

The objective of this second evaluation test was to assess the fluency of MT. According to the Linguistic Data Consortium⁴, fluency is the degree to which a translation "is well-formed grammatically, contains correct spellings, adheres to common use of terms, titles and names, is intuitively acceptable and can be sensibly interpreted by a native speaker." According to TAUS DQF, the degree of sentence fluency should be scored as follows:

1. Incomprehensible: Refers to a very poorly written text that is impossible to understand.
2. Disfluent: Refers to a text that is poorly written and difficult to understand.
3. Good: Refers to a smoothly flowing text even when a number of minor errors are present.
4. Flawless: Refers to a perfectly flowing text with no errors.

Adequacy

In this third test, the aim was to assess the adequacy of MT. According to the Linguistic Data Consortium, adequacy is "how much of the meaning expressed in the gold-standard translation or the source is also expressed in the target translation." According to TAUS DQF, the degree of adequacy of sentences should be scored as follows:

1. None: None of the meaning in the source is contained in the translation.
2. Little: Fragments of the meaning in the source are contained in the translation.
3. Most: Almost all the meaning in the source is contained in the translation.
4. Everything: All the meaning in the source is contained in the translation, no more, no less.

After the MT Ranking and Fluency and Adequacy tests were performed, the first main factor of the MT human evaluation was obtained: the quality of the MT systems. The MT Ranking evaluation indicated which MT system was best, while the Fluency and Adequacy evaluations gave feedback on precision and fluency, all of which illustrated whether there was a clear winner or if there was an engine that was better in fluency but worse in adequacy or vice versa.

For the Adequacy and Fluency tests, the evaluation was performed by the author, who had a degree in Translation and Interpreting, a Master's degree in Specialised Translation, and +5 years of professional experience as a freelance translator in

⁴Please see <https://www ldc.upenn.edu>, last accessed on 21st September 2021.

specialised domains. It would have been helpful to recruit a professional translator other than the author, but this was impossible due to budget constraints.

Productivity

The second big factor to be studied in our human evaluation was productivity, and we also used the TAUS DQF to this end. Evaluators saw a sentence in the source language, as well as the MT raw output of one of the MT systems and had to post-edit the machine translation to achieve a professional human-quality translation. This test provided us with the following data:

- Editing time: the time it took for the evaluators to do the post-editing, as well as the average number of words that could be post-edited per hour.
- Editing distance: a number that indicated the modifications that the evaluator made during the post-editing process. The more modifications evaluators made to the raw output, the higher the number. If no modifications were made, the editing distance was 0.

Post-editing guidelines, which indicated the instructions and criteria to be followed in the post-editing process, were also provided for the productivity test. In addition, the MT systems were analysed in two study groups, as the objectives were different:

1. On the one hand, six participants of the MT ranking test evaluated Softcatalà's Translator and Google Translate (both NMT engines). From this comparison, we would obtain results on which engine was most useful in the English-Catalan language pair.
2. On the other hand, six members of Softcatalà (volunteers) evaluated Softcatalà's Translator and Apertium (both open-source engines). From this comparison, we would know which engine Softcatalà members should use for the translation and localisation of the projects in which they collaborate.

	T1 System 1	T2 System 1	T1 System 2	T2 System 2
PE1	x			x
PE2		x	x	
PE3	x			x
PE4		x	x	
PE5	x			
PE6		x		x

Table 1. Overview of the text assignment methodology

As this test was more convoluted and the results were more complex to analyse, the number of evaluators was reduced to 6 per group of study. By having to compare the post-editing of two different MT systems, evaluators could not post-edit the same text twice. To avoid the learning effect of post-editors working with the same texts, we created two different texts (T1 and T2). These texts were assigned as follows, so all translators worked under all conditions and with different texts (see Table 1)

3.4. Selection and preparation of the texts

Text types and domains are crucial in the training process or the evaluation of an MT system. In this study, the text evaluated came from Home Assistant⁵, a virtual assistant for smart homes. The reasons for choosing this text were that Home Assistant was an open-source product, and segments and text could be obtained easily. Furthermore, software localisation is one of the most common types of translation in today's language services industry (EUATC, 2020). Thus, the aim of this study was to obtain results that were suitable and valid for the current context of the industry.

The text was downloaded from Home Assistant's GitHub repository, and 200 segments were randomly selected for two different texts (T1 and T2). T1 contained 1,006 words, and T2 had 1,065. In this selection, long segments were randomly mixed with short segments (as the latter are very frequent in software localisation), as well as with placeholders and variables. After selecting the text samples, all the segments were machine translated, and the raw MT output was pre-processed as per TAUS DQF's platform requirements. The final files were then uploaded into the TAUS DQF tool, and evaluators received an automatic email with a link to the corresponding test.

4. Analysis of the results

4.1. MT Ranking

In this test, 100 segments were evaluated. Figure 1 shows which MT system was rated as the best MT, i.e., which proposal would need fewer edits to obtain a segment with professional human quality. The percentages in the graph indicate how many times the evaluators assigned Score 1 (the best MT proposal) to a specific engine. Softcatalà's Translator was rated as the best engine (40.6% of the time) compared to Google Translate (39.7%) and Apertium (19.7%). Although the distance between the first and the second engine is minimal (only 0.9 percentage points), if we look at the data in depth, 10 out of the 11 evaluators rated Softcatalà's Translator as the best MT system. The only evaluator who did not think this rated both engines as equal.

⁵Please see <https://www.home-assistant.io>, last accessed on 21st September 2021.

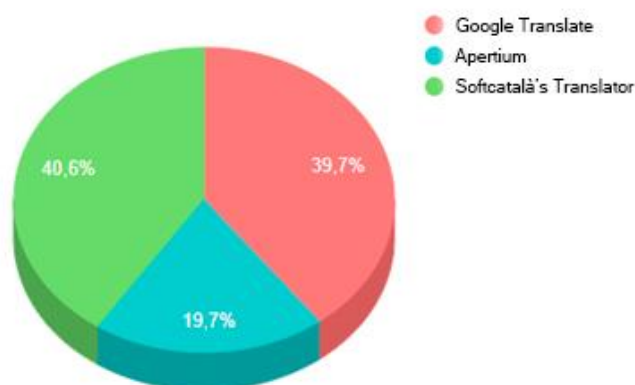


Figure 1. Overall MT Ranking results: percentage of times an engine has received Score 1 (the best translation proposal)

If we look at the data in another way and create a table per MT system and indicate the percentage of times the MT system has received each of the scores, whether 1, 2 or 3, we obtain the following graph.

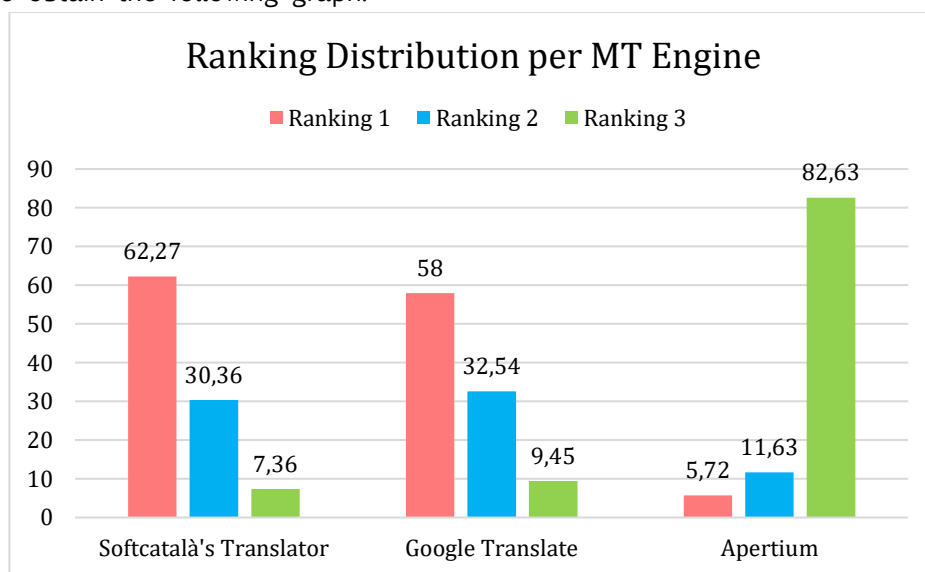


Figure 2. MT Ranking results per MT engine: distribution of rankings

Figure 2 shows that Softcatalà's Translator was not only the MT system regarded as the best engine most often (Score 1), but also the MT system with the fewest ratings as the worst engine (Score 3). According to the perception of the evaluators and almost unanimously, Softcatalà's Translator was the MT engine that offered the best translations from the MT systems evaluated in the English-Catalan language pair. It is worth stressing, however, that both top-performing MT systems obtained excellent results.

4.2. Fluency and Adequacy

In these tests, the translation proposals were also anonymous, and the evaluator did not know which engine provided each translation proposal. Figure 3 shows the results of the fluency evaluation.

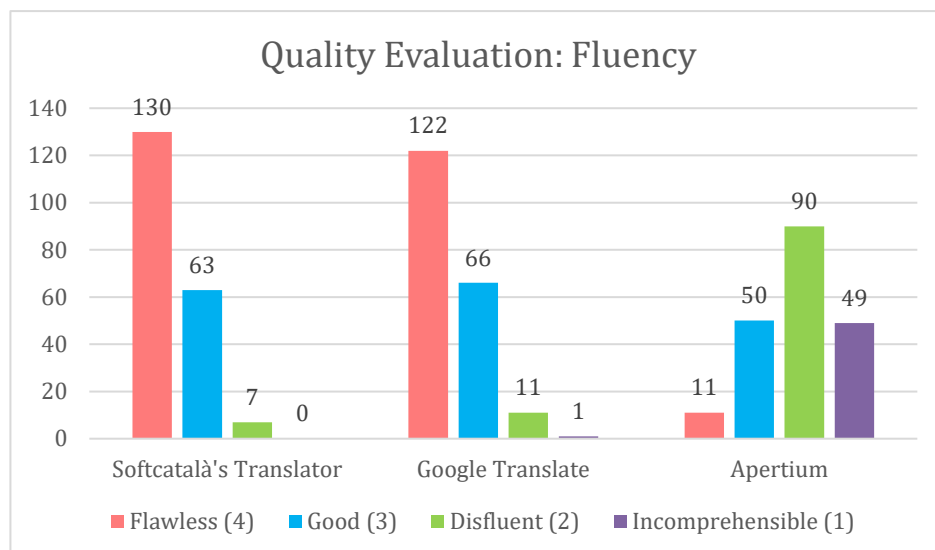


Figure 3. Results of the fluency evaluation

Softcatalà's Translator obtained the best score, as 130 segments had perfect fluency and 63 segments had good fluency. This meant 96.5% of Softcatalà's segments offered perfect or good fluency, allowing the text to be read naturally and in accordance with the rules of the target language. Google Translate obtained very similar results, although slightly lower, with 122 segments rated with perfect fluency and 66 segments with good fluency. This meant that Google Translate achieved perfect or good fluency in 94% of the cases. Apertium, on the other hand, had very different results, and only 30.5% of the segments were rated among the top two fluency categories.

Although the MT Ranking classification was relative, this fluency evaluation allowed us to see that Softcatalà's Translator and Google Translate offered more fluent translations than Apertium. While Softcatalà's Translator and Google Translate obtained 0 and 1 segments with the score 1 (Incomprehensible), Apertium received this score in 49 segments. This reinforced the initial hypothesis of the paper, since we assumed that NMT engines would offer better fluency than the previous MT paradigms, in line with other studies on NMT (Castilho et al., 2018, 2017a).

Figure 4 shows the adequacy results, which reflect a similar reality to that of fluency: there is a big difference in quality between Softcatalà's Translator and Google Translate, on the one hand, and Apertium, on the other hand.

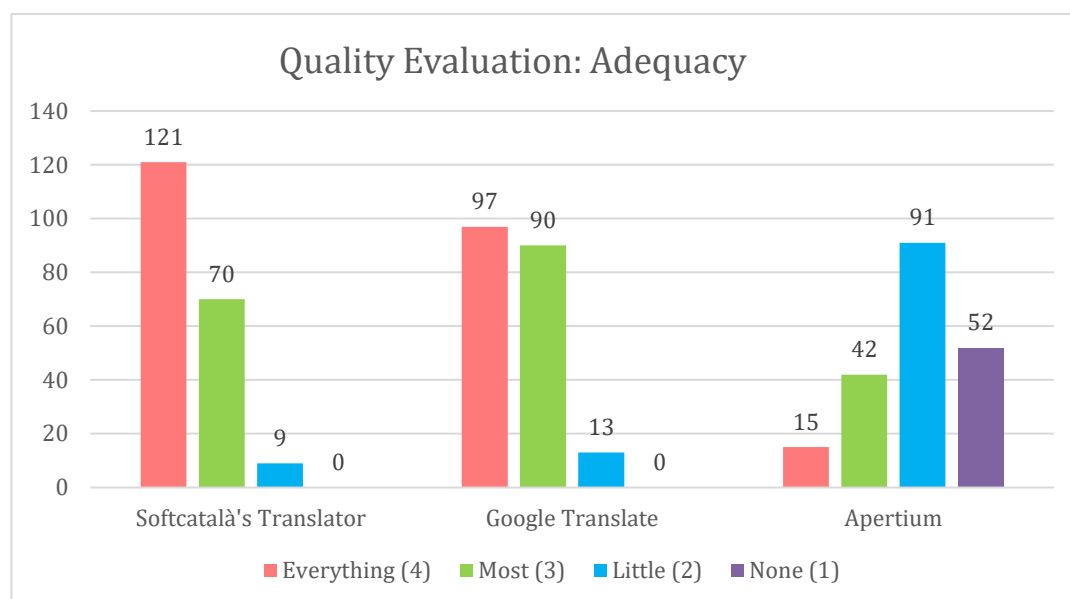


Figure 4. Results of the adequacy evaluation

Softcatalà's Translator and Google Translate were ranked 191 and 187 segments out of 200 in the two highest adequacy categories. However, if evaluated in more detail, Softcatalà's Translator had 121 segments with score 4 (Everything), while Google Translate had only 97; regarding score 3 (Most); both systems were assigned this score in 70 vs 90 segments, respectively. This indicated that, although both MT engines offered very accurate translations, Softcatalà's Translator was more precise than Google Translate. As for Apertium, 58% of the total segments obtained the two lowest scores (Little and None), which indicated that the raw translations proposed by Apertium did not reflect the meaning of the original text and adequacy was far behind the other two MT systems in this study.

4.3. Productivity

As to productivity, evaluators performed the post-editing tasks, and we obtained the post-editing time of each segment in milliseconds, as well as the editing distance (the more additions, deletions or changes to the raw translation proposed by the MT engine, the higher the number). The post-editing times of all participants were summed for each engine, and the mean (in milliseconds and seconds) and the standard deviation (SD) were calculated to statistically analyse the average post-editing time per person. Subsequently, we analysed the statistical significance of the results to study whether the results obtained could have been produced by chance, or whether there was a clear statistical significance. If the p-value was less than 0.05, we could say that the samples were statistically significant. We also followed the same process in terms of edit distance. Additionally, to further analyse the results, segments were split into three groups according to their length in the source language: from 1 to 5 words (44 segments); from 6 to 15 words (113 segments); and with 16 or >16 words (43 segments). Then, data were statistically analysed in the same way as in the general analysis: calculation of the mean in milliseconds and seconds, calculation of the SD and statistical significance test.

Hereafter, PE time is shown in seconds for ease of understanding. As there were two different groups of study, results were also divided in two different sections.

4.3.1. Study Group 1 (Softcatalà's Translator-Google Translate)

The general productivity results obtained from the first group of evaluators is shown in Table 2.

	Softcatalà's Translator		Google Translate	
	Mean	SD	Mean	SD
PE time (s)	3909.07	21.61	4131.64	18.89
Edit distance (segment-level)	9.79	13.87	10.35	13.92
Student's T (PE time)	0.19 (> 0.05)			
Student's T (edit distance)	0.42 (> 0.05)			

Table 2. Overall productivity results: study group 1

The overall post-editing time was shorter in Softcatalà's Translator than in Google Translate (a difference of 222.56 seconds; 5.69% less). The overall editing distance, analysed at the segment level, was lower in Softcatalà's Translator (9.79) than in Google Translate (10.35), with a total difference of 0.56 points. Student's t-test yielded values greater than 0.05, so we could not affirm that the results were statistically significant.

		1-5 words		6-15 words		16 or >16 words	
		Mean	SD	Mean	SD	Mean	SD
PE time (s)	Softcatalà	8.15	9.37	18.44	17.92	34.08	29.97
	Google	9.41	8.84	20.08	17.60	33.67	21.69
Student's T		0.22 (> 0.05)		0.12 (> 0.05)		0.87 (> 0.05)	

Table 3. Specific productivity results: post-editing time, study group 1

When analysing the length of the segments (in words), as can be seen in Table 3, the post-editing time of Softcatalà's Translator's was shorter in the segments from 1 to 5 words (8.15 seconds vs. 9.41 seconds) and also in those segments from 6 to 15

words (18.44 seconds vs. 20.08 seconds). These results were the average post-editing time by segment length. However, for long segments of 16 or more than 16 words, post-editing time was shorter with Google Translate (34.08 seconds vs. 33.67 seconds). Our statistical analyses also indicated that these differences were not statistically significant either.

		1-5 words		6-15 words		16 or >16 words	
		Mean	SD	Mean	SD	Mean	SD
Edit distance (segment)	Softcatalà	5.34	13.68	11.53	14.81	9.79	10.05
	Google	12.22	19.92	9.31	11.96	11.20	10.76

Levene's test	0.001 (< 0.05)	0.006 (< 0.05)	0.25 (> 0.05)
Student's T	0.0008 (< 0.05)	0.007 (< 0.05)	0.19 (> 0.05)

Table 4. Specific productivity results: edit distance, study group 1

According to the results in Table 4, the edit distance at the segment level was lower in the segments from 1 to 5 words in Softcatalà's Translator (5.34 vs. 12.22). Yet, in medium-length segments, the edit distance was lower with Google Translate (9.313 vs. 11.53). In these two cases, the p-value was under 0.05, and we could therefore claim statistical significance. On the other hand, for segments with 16 or more than 16 words, the editing distance was again lower in Softcatalà's Translator (9.79 vs. 11.20), but our statistical tests showed no statistical significance.

To summarise the previous tables and simplify the analysis of this first group of study, Table 5 shows an overall summary of the results. The asterisk symbol (*) indicates that results were statistically significant.

	PE time	Edit distance
1-5 words	Softcatalà < Google	Softcatalà < Google*
5-15 words	Softcatalà < Google	Softcatalà > Google*
16 or >16 words	Softcatalà > Google	Softcatalà < Google

Table 5. Summary of productivity results: post-editing time and editing distance, study group 1

As a global summary of the results of the first group of study, we can state that the average post-editing time was shorter with Softcatalà's Translator than with Google Translate. When speaking about the different types of segments, short- and medium-length segments' post-editing time was also shorter in Softcatalà's Translator, although it was bigger in longer segments. However, most of these results were not statistically significant, and it would be advisable to

increase the samples and re-test. On the other hand, as far as the editing distance is concerned, we can state that the edit distance was higher for Google Translate in segments of 1 to 5 words, but slightly lower in segments of 5 to 15 words. These results were statistically significant. In segments of 16 or more than 16 words, the editing distance was smaller for Softcatalà's Translator, although it would be advisable to increase the samples and re-test to obtain statistically significant results in this type of segments.

4.3.2. Study Group 2 (Softcatalà's Translator-Apertium)

For the second study group, Softcatalà's Translator was compared against Apertium by following the same methodology used in the analysis of the first study group.

	Softcatalà's Translator		Apertium	
	Mean	SD	Mean	SD
PE time (s)	1859.51	12.57	3743.41	21.84
Edit distance (segment)	6.81	12.68	24.85	20.20
Student's T (PE time)	2.42E-39 (< 0.05)			
Student's T (edit distance)	1.171E-135 (< 0.05)			

Table 6. Overall productivity results: study group 2

According to Table 6, we can see that the overall post-editing time was significantly shorter in Softcatalà's Translator than in Apertium (a difference of 1883.89 seconds; 101.31% shorter). As for the overall editing distance, we can also observe that the editing distance at the segment level was substantially smaller in Softcatalà's Translator (6.81) than in Apertium (24.85). The statistical analyses showed that these results were statistically significant. We can therefore state that post-editing time and editing distance were much shorter in Softcatalà's Translator than in Apertium. To analyse these data in more depth, Table 7 shows these results in terms of segment-length.

		1-5 words		6-15 words		16 or >16 words	
		Mean	SD	Mean	SD	Mean	SD
PE time (s)	Softcatalà	5.95	6.14	14.18	15.80	25.83	16.21
	Apertium	14.11	14.88	28.64	18.55	55.70	28.09
Student's T		4.49E-08 (< 0.05)		2.434E-26 (< 0.05)		3.8408E-19 (< 0.05)	

Table 7. Specific productivity results: post-editing time, study group 2

Post-editing time of Softcatalà's Translator was much lower in all the segment-lengths in comparison with Apertium: segments from 1 to 5 words (5.95 seconds vs. 14.11 seconds; 136.91% lower), from 6 to 15 words (14.18 seconds vs. 28.64 seconds; 101.89% lower), as well as in those with 16 or more than 16 words (25.83 seconds vs. 55.70 seconds; 115.58% lower). Differences were statistically significant in all segment-lengths.

		1-5 words		6-15 words		16 or >16 words	
		Mean	SD	Mean	SD	Mean	SD
Edit distance (segment)	Softcatalà	6.21	11.85	10.73	13.40	10.11	8.70
	Apertium	40.65	29.53	37.76	18.13	36.15	13.39

Student's T	9.709E-26 (< 0.05)	9.535E-83 (< 0.05)	3.937E-44 (<0.05)
-------------	--------------------	--------------------	-------------------

Table 8. Specific productivity results: edit distance, study group 2

Table 8 shows that edit distance was also substantially lower in all segment-lengths: in short segments from 1 to 5 words (6.21 vs. 40.65), in medium-length segments from 6 to 15 words (10.73 vs. 37.76), but also in long segments with 16 or more than 16 words (10.11 vs. 36.15). These huge differences were statistically significant and suggested that edit distance was significantly lower when using Softcatalà's Translator instead of Apertium.

	PE time	Edit distance
1-5 words	Softcatalà < Apertium*	Softcatalà < Apertium*
5-15 words	Softcatalà < Apertium*	Softcatalà < Apertium*
16 or >16 words	Softcatalà < Apertium*	Softcatalà < Apertium*

Table 9. Summary of productivity results: post-editing time and edit distance, study group 2

As in the first group of study, Table 9 gathers and simplifies the analysis of this second study group. Unlike the first study group, where there was one engine that slightly outperformed the other in most aspects (although some results were not statistically significant), in this second group of study we could clearly see that there was a big difference between the engines analysed. Results indicated that Softcatalà's Translator reduced the post-editing times and the editing distances in comparison with Apertium in all segment-lengths, and these differences were statistically significant.

5. Conclusions

In this paper, we stressed the importance of hiring professional evaluators and using strict controlled procedures to ensure that results are valid, and that there are no flaws in the methodology used. To this end, a series of human evaluations were performed to assess the MT engines analysed in this paper and to meet the objectives of the study. Different guidelines were created to help evaluators homogenize their criteria when performing the evaluation and reduce bias.

With regard to the first objective (To analyse which MT method and system of choice offered higher translation quality), translation quality was evaluated with the MT Ranking, Fluency and Adequacy tests of TAUS DQF.

In the MT Ranking test, evaluators decided almost unanimously that Softcatalà's Translator was the best engine in the English-Catalan language pair, since, according to the perception of 10 out of the 11 evaluators, Softcatalà's Translator achieved the best ranking in most segments. The person who did not rate Softcatalà's Translator as the best MT system thought that Softcatalà's Translator and Google Translate were equally good. In terms of the Fluency test, Softcatalà's Translator and Google Translate obtained similar results, with 96.5% and 94% of their segments with a score of perfect or good fluency. Apertium, on the other hand, was relegated to last position with perfect or good fluency scores in only 30.5% of its segments. The remainder, 69.5% of the total segments, were rated poor or incomprehensible Fluency scores. As far as the Adequacy test is concerned, Softcatalà's Translator continued to be in first position, slightly distancing itself from Google Translate, although both obtained excellent results. Apertium, once again, lagged far behind in terms of adequacy, obtaining feeble scores in comparison with the other two engines. After these three tests, though Softcatalà's Translator was the best MT engine according to the perceptions of the evaluators, differences between the top 2 MT systems were almost identical, and we can conclude that both Softcatalà's Translator and Google Translate are great MT options for the English-Catalan language pair. It is worth stressing, however, that Softcatalà's Translator is an open-source MT system, the data used to train the system can be accessed easily and can be used to create other MT systems, and privacy is respected, in contrast with the case of proprietary MT engines (Moorkens, 2022).

Regarding the second objective (2. To find out which MT engine offered the best post-editing performance), an analysis of productivity was carried out by studying the post-editing tasks of two different groups of study that worked on two different texts.

On the one hand, when analysing the results of the first study group (working with Softcatalà's Translator and Google Translate), we could affirm that the post-editing time and editing distance was slightly shorter for Softcatalà's Translator in comparison with Google Translate. In other words, more text was translated in less time when using Softcatalà's Translator and, furthermore, fewer modifications (additions, edits or omissions) had to be made to the raw output. However, in most cases, the results were not statistically significant, and it would be advisable to increase the size of the evaluations (by increasing the number of segments or the number of evaluators). On the other hand, when analysing the results of the second study group (working with Softcatalà

and Apertium), we could state categorically that post-editing time and edit distance were much shorter for Softcatalà's Translator in comparison with Apertium. These results were statistically significant in all the assumptions analysed. These results backed our hypothesis that NMT engines were more accurate and fluent than rule-based MT engines and, although NMT errors may be more difficult to spot due to the fluency of the raw output, post-editing NMT output is still more productive than post-editing rule-based MT output.

Consequently, and after analysing the data obtained as a whole, we could state that both Softcatalà's Translator and Google Translate offered excellent translation quality, proving to be interesting MT options in a workflow whose objective was dissemination in the English-Catalan language pair. Depending on the confidentiality of the documentation to translate, we may prefer to opt for Softcatalà's Translator, instead of opting for Google's proprietary MT system. It is also worth stressing that the translations analysed were produced in June 2020, and that both Softcatalà and Google may have updated and re-trained their systems, which would give different results in the future. Nevertheless, these results are an interesting indicator for anyone interested in MT in the English-Catalan combination. Besides the answers that this study provides, the results opened up a series of questions that could be interesting to address in future studies:

- Are these results the same in other fields? Since we only tested these engines in software localisation, it would be interesting to evaluate sectors such as the legal or medical domains.
- The data obtained could be used to analyse other very compelling aspects. What results would we obtain if we also separate the segments not only by their length, but also by their quality, in accordance with the results of fluency and adequacy? What changes did post-editors make in the segments of higher quality? And in those of lower quality? This way, post-editors' behaviour according to segment quality could also be analysed.
- Furthermore, to improve the quality of Softcatalà's Translator, is it more useful to increase the number of texts to train the engine or to improve the quality of the existing texts? What should be prioritised, the volume or the quality of the texts when re-training the engine?

Despite new research questions that have arisen, the present work has clarified that both Softcatalà's Translator and Google Translate are, at the date of writing, the best MT engines publicly available for the English-Catalan combination in software localisation. Yet, bearing in mind that Softcatalà's Translator is open-source, it encompasses benefits in terms of price, character limitation, translation of large documents and confidentiality, which may be at stake in a proprietary NMT system.

6. References

Ahrenberg, L. (2017). Comparing Machine Translation and Human Translation: A Case Study. In: The Proceedings of the Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT): Varna, Bulgaria, september 2017. Shoumen, Bulgaria: Incoma, pp. 21-28. <https://doi.org/10.26615/978-954-452-042-7_003>. [Accessed: 20221209].

- Barrault, L.; Biesialska, M.; Bojar, O.; Costa-jussà, M. R.; Federmann, C.; Graham, Y.; et al. (2020a). Findings of the 2020 Conference on Machine Translation (WMT20). In: Proceedings of the Fifth Conference on Machine Translation (WMT), online, November 19-20, 2020. Stroudsburg, PA: Association for Computational Linguistics, pp. 1-55. <<https://aclanthology.org/2020.wmt-1.1.pdf>>. [Accessed: 20221209].
- Barrault, L.; Bojar, O.; Bougares, F.; Chatterjee, R.; Costa-jussà, M. R.; Federmann, C.; et al. (eds.) (2020b). Proceedings of the Fifth Conference on Machine Translation (WMT), online, November 19-20, 2020. Stroudsburg, PA: Association for Computational Linguistics. <<https://aclanthology.org/volumes/2020.wmt-1/>>. [Accessed: 20221209].
- Barrault, L.; Bojar, O.; Bougares, F.; Chatterjee, R.; Costa-jussa, M. R.; Federmann, C.; et al. (eds.). (2021). Proceedings of the Sixth Conference on Machine Translation, online, November 10-11, 2021. Stroudsburg, PA: Association for Computational Linguistics. <<https://aclanthology.org/2021.wmt-1.0.pdf>>. [Accessed: 20221209].
- Bentivogli, L.; Bisazza, A.; Cettolo, M.; Federico, M. (2016). Neural versus Phrase-Based Machine Translation Quality: a Case Study. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, Texas November 1-5 2016. Stroudsburg, PA: Association for Computational Linguistics, pp. 257-267. <<https://doi.org/10.18653/v1/D16-1025>>. [Accessed: 20221209].
- Bojar, O.; Chatterjee, R.; Federmann, C.; Graham, Y.; Haddow, B.; Huck, M.; et al. (2016). Findings of the 2016 Conference on Machine Translation. In: Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers: Berlin. Germany, August 11-12, 2016. Stroudsburg, PA: Association for Computational Linguistics, pp. 131-198. <<https://doi.org/10.18653/v1/W16-2301>>. [Accessed: 20221209].
- Briva-Iglesias, V., (2021). Traducción humana vs. traducción automática: análisis contrastivo e implicaciones para la aplicación de la traducción automática en traducción jurídica. *Mutatis mutandis: revista latinoamericana de traducción*, v. 14, n. 2, pp. 571-600. <<https://doi.org/10.17533/udea.mut.v14n2a14>>. [Accessed: 20221209].
- Callison-Burch, C.; Fordyce, C.; Koehn, P.; Monz, C.; Schroeder, J. (2007). (Meta-) evaluation of machine translation. In: StatMT'07: Proceedings of the Second Workshop on Statistical Machine Translation: Prague, June 2007. Stroudsburg, PA: Association for Computational Linguistics, pp. 136-158. <<https://doi.org/10.3115/1626355.1626373>>. [Accessed: 20221209].
- Castilho, S.; Doherty, S.; Gaspari, F.; Moorkens, J. (2018). Approaches to Human and Machine Translation Quality Assessment. In: Moorkens, J.; Castilho, S.; Gaspari, F.; Doherty, S. (eds.). *Translation Quality Assessment: From Principles to Practice*. Cham: Springer. (Machine Translation: Technologies and Applications; 1), pp. 9-38. <https://doi.org/10.1007/978-3-319-91241-7_2>. [Accessed: 20221209].
- Castilho, S.; Moorkens, J.; Gaspari, F.; Calixto, I.; Tinsley, J., Way, A. (2017a). Is Neural Machine Translation the New State of the Art? *The Prague Bulletin of Mathematical*

- Linguistics, n. 108 (June), pp. <109–120. <https://doi.org/10.1515/pralin-2017-0013>>. [Accessed: 20221209].
- Castilho, S.; Moorkens, J.; Gaspari, F.; Sennrich, R.; Sosoni, V.; Georgakopoulou, P.; et al. (2017b). A Comparative Quality Evaluation of PBSMT and NMT using Professional Translators. In: Kurohashi, S.; Fund, P. (ed.). Proceedings of Machine Translation Summit XVI: Research Track, pp. 116-131. <https://aclanthology.org/2017.mtsummit-papers.10/>>. [Accessed: 20221209].
- Cronin, M. (2012). Translation in the Digital Age. 1st ed. Milton Park, Abigdon [etc.]: Routledge. <https://doi.org/10.4324/9780203073599>>. [Accessed: 20221209].
- EUATC (2020). 2020 European Language Industry Survey launched. <https://euatc.org/industry-surveys/2020-language-industry-survey-launched/>>. [Accessed: 20221209].
- Faes, F. (2016). Disruption Turns 10 as Google Translate Comes of Age. Slator. <https://slator.com/disruption-turns-10-google-translate-comes-age/>>. [Accessed: 20221209].
- Forcada, M. L.; Ginestí-Rosell, M.; Nordfalk, J.; O'Regan, J.; Ortiz-Rojas, S.; Pérez-Ortiz, J. A.; et al. (2011). Apertium: a free/open-source platform for rule-based machine translation. Machine Translation, n. 25, pp. 127-144. <https://doi.org/10.1007/s10590-011-9090-0>>. [Accessed: 20221209].
- Freitag, M.; Foster, G.; Grangier, D.; Ratnakar, V.; Tan, Q.; Macherey, W. (2021). Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. Transactions of the Association for Computational Linguistics, n. 9, pp. 1460-1474. https://doi.org/10.1162/tacl_a_00437>. [Accessed: 20221209].
- Freitag, M.; Grangier, D.; Caswell, I. (2020). BLEU might be Guilty but References are not Innocent. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), online, pp. 61-71. <https://doi.org/10.18653/v1/2020.emnlp-main.5>>. [Accessed: 20221209].
- Görög, A. (2014). Quantifying and benchmarking quality: the TAUS Dynamic Quality Framework. Revista Tradumàtica: tecnologies de la traducció, n. 12, pp. 443-454. <https://doi.org/10.5565/rev/tradumatica.66>>. [Accessed: 20221209].
- Graham, Y.; Baldwin, T.; Moffat, A.; Zobel, J. (2013). Continuous Measurement Scales in Human Evaluation of Machine Translation. In: Proceedings of the 7th Linguistics Annotation Workshop and Interoperability with Discourse: Sofia, Bulgaria, August 8-9 2013. Stroudsburg, PA: Association for Computational Linguistics, pp. 33-41. <https://aclanthology.org/W13-2305.pdf>>. [Accessed: 20221209].
- Kocmi, T.; Federmann, C.; Grundkiewicz, R.; Junczys-Dowmunt, M.; Matsushita, H.; Menezes, A. (2021). To Ship or Not to Ship: An Extensive Evaluation of Automatic Metrics for Machine Translation. In: Proceedings of the Sixth Conference on Machine Translation (WMT), November 10-11, 2021. Stroudsburg, PA: Association for Computational Linguistics, pp. 478-494. <https://aclanthology.org/2021.wmt-1.57.pdf>>. [Accessed: 20221209].

- Koehn, P.; Monz, C. (2006). Manual and automatic evaluation of machine translation between European languages. In: StatMT'06: Proceedings of the Workshop on Statistical Machine Translation, New York City, June 2006. Stroudsburg, PA: Association for Computational Linguistics, pp. 102-121.
<<https://doi.org/10.3115/1654650.1654666>>. [Accessed: 20221209].
- Läubli, S.; Castilho, S.; Neubig, G.; Sennrich, R.; Shen, Q.; Toral, A. (2020). A Set of Recommendations for Assessing Human-Machine Parity. In: Language Translation. Journal of Artificial Intelligence Research, v. 67, pp. 653-672.
<<https://doi.org/10.1613/jair.1.11371>>. [Accessed: 20221209].
- Läubli, S.; Sennrich, R.; Volk, M. (2018). Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31-November 4, 2018. Stroudsburg, PA: Association for Computational Linguistics, pp. 4791-4796. <<https://doi.org/10.18653/v1/D18-1512>>. [Accessed: 20221209].
- López-Pereira, A. (2019). Traducción automática neuronal y traducción automática estadística: percepción y productividad. Revista Tradumàtica: tecnologies de la traducció, n. 17, pp. 1-19. <<https://doi.org/10.5565/rev/tradumatica.235>>. [Accessed: 20221209]
- Martín-Mor. A.; Piqué, R.; Sánchez-Gijón, P. (2016). Tradumàtica: Tecnologies de la traducció. EUMO, Vic.
- Moorkens, J. (2022). Ethics and machine translation. In: Kenny, Dorothy (ed.). Machine translation for everyone: Empowering users in the age of artificial intelligence. Berlin: Language Science Press. (Translation and Multilingual Natural Language Processing; 18), pp. 121-140.
- Moorkens, J.; Castilho, S.; Gaspari, F.; Doherty, S. (eds.) (2018). Translation Quality Assessment: From Principles to Practice. Cham: Springer. (Machine Translation: Technologies and Applications; 1). <<https://doi.org/10.1007/978-3-319-91241-7>>. [Accessed: 20221209].
- O'Brien, S. (2012). Towards a dynamic quality evaluation model for translation. The Journal of Specialised Translation, n. 17 (January), pp. 55-77.
<https://www.jostrans.org/issue17/art_obrien.pdf>. [Accessed: 20221209].
- Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.-J. (2001). BLEU: a method for automatic evaluation of machine translation. In: ACL'02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, pp. 311-318.
<<https://doi.org/10.3115/1073083.1073135>>. [Accessed: 20221209].
- Pitman, J. (2021). Google Translate: One billion installs, one billion stories. Google.
<<https://blog.google/products/translate/one-billion-installs/>>. [Accessed: 20221209].
- Popović, M. (2015). chrF: character n-gram F-score for automatic MT evaluation. In: Proceedings of the Tenth Workshop on Statistical Machine Translation. Stroudsburg,

- PA: Association for Computational Linguistics, pp. 392–395.
<<https://doi.org/10.18653/v1/W15-3049>>. [Accessed: 20221209].
- Rei, R.; Stewart, C.; Farinha, A. C.; Lavie, A. (2020). COMET: A Neural Framework for MT Evaluation. In: Proceedings of the 2020 Conference on empirical Methods in Natural Language Processing (EMNLP), online. Stroudsburg, PA: Association for Computational Linguistics, pp. 2685–2702. <<https://aclanthology.org/2020.emnlp-main.213/>>. [Accessed: 20221209].
- Snover, M.; Dorr, B.; Schwartz, R.; Micciulla, L.; Makhoul, J. (2016). A Study of Translation Edit Rate with Targeted Human Annotation. In: Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers. Association for Machine Translation in the Americas, pp. 223–231. <<https://aclanthology.org/2006.amta-papers.25.pdf>>. [Accessed: 20221209].
- Snover, M.; Madnani, N.; Dorr, B.; Schwartz, R. (2009). TER-Plus: paraphrase, semantic, and alignment enhancements to Translation Edit Rate. Machine Translation, n. 23, pp. 117–127. <<https://doi.org/10.1007/s10590-009-9062-9>>. [Accessed: 20221209].
- Tillmann, C.; Vogel, S.; Ney, H.; Zubiaga, A.; Sawaf, H. (1997). Accelerated DP based search for statistical translation. <https://www.isca-speech.org/archive_v0/archive_papers/eurospeech_1997/e97_2667.pdf>. [Accessed: 20221209].
- Toral, A. (2020). Reassessing Claims of Human Parity and Super-Human Performance in Machine Translation at WMT 2019. In: Proceedings of the 22nd Annual Conference of the European Association for Machine Translation. European Association for Machine Translation, pp. 185–194. <<https://aclanthology.org/2020.eamt-1.20.pdf>>. [Accessed: 20221209].
- Wu, Y.; Schuster, M.; Chen, Z.; Le, Q. V.; Norouzi, M.; Macherey, W.; et al. (2016). Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. Cornell University. <<https://doi.org/10.48550/arXiv.1609.08144>>. [Accessed: 20221209].
- Ye-Yi Wang; Acero, A.; Chelba, C. (2003). Is word error rate a good indicator for spoken language understanding accuracy. In: 2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No.03EX721), pp. 577–582. <<https://doi.org/10.1109/ASRU.2003.1318504>>. [Accessed: 20221209].

Annex 1. Guidelines for the MT Ranking evaluation

Objective

The aim of this document is to clearly define guidelines and instructions for the human evaluation of machine translation (MT). In this evaluation task, you will perform a relative assessment of the translation proposals of different MT engines.

Description of the test

1. You will receive an email with the assessment system link and brief instructions. By clicking on the link, you will go to the page where the assessment is made.
2. On this page, click the "Start" button to start.
3. You will evaluate 100 segments. In the user interface you will see "Source" with the English segment and "Target" with three Catalan sentences. Each of these sentences comes from an anonymous MT engine. See the image below.

Rànquing de TA (Rank Comparison)

Source (English (United Kingdom))

Start

Current This entity does not have a unique ID, therefore its settings cannot be managed from the UI.

Next The {platform} integration is not loaded.

Target (Catalan)

0 Aquesta entitat no té un ID únic, per tant la seva configuració no es pot gestionar des de la IU.

0 Aquesta entitat no té un ID únic, per tant, la seva configuració no es pot gestionar des de la interfície d'interès.

0 Aquesta entitat no té un únic ID, per tant no es poden abastar els seus paràmetres des del UI. [\(Info\)](#)

Comments

Characters left: 500

1. You must read all translations and make a relative evaluation of these engines. What does that mean? You must assign the following scores:

1	This is the best translation of the three proposed
2	This is the second best translation of the three proposed
3	This is the worst translation of the three proposed

2. In case of draw with two translation proposals, you can assign a number 1 score to the two best options, and assign number 3 to the worst sentence, or vice versa.
3. If you want, you can use the "Comments" field to make clarifications (optional).
4. You can return to previous segments by clicking the "Previous" button.
5. You can pause the evaluation by clicking "Home".
6. If you haven't logged out of your browser, click "Continue" to resume your evaluation.
7. If you've logged out of your browser, click the link to the evaluation in the email you received, and then click "Continue" to resume the evaluation.

Annex 2. Guidelines for the Fluency and Adequacy evaluations

Objectives

The aim of this document is to clearly define guidelines and instructions for the human evaluation of machine translation (MT). In this evaluation task, you will assess the fluency and adequacy offered by different raw translations of different MT engines.

Description of the test

1. You will receive an email with the assessment system link and brief instructions. By clicking on the link, you will go to the page where the assessment is made.
2. On this page, click the "Start" button to start.
3. You will evaluate 100 segments. In the user interface you will see "Source" with the English segment and "Target" with three Catalan sentences. Each of these sentences comes from an anonymous MT engine. See the image below.

Precisió i fluïdesa S2, TA2

The screenshot shows a web interface for evaluating machine translation. It is divided into four main sections:

- Source (English (United Kingdom))**: Contains a "Start" button, a "Current" segment with the text "This service is run by our partner, a company founded by the founders of Home Assistant and Hass.io.", and a "Next" button with the text "Go to the integrations page."
- Target (Catalan)**: Contains a "Start" button, a "Current" segment with the text "Aquest servei el gestiona el nostre soci, una empresa fundada pels fundadors de Home Assistant i Hass.io.", and a "Next" button with the text "Vés a la pàgina d'integracions."
- Fluency:**: Contains four radio button options: "Incomprehensible", "Disfluent", "Good", and "Flawless". A "(More Info)" link is on the right.
- Adequacy:**: Contains four radio button options: "None", "Little", "Most", and "Everything". A "(More Info)" link is on the right.

1. You have to read the MT proposal and evaluate its fluency. What does that mean? You must assign one of the following scores:

Incomprehensible	The sentence is badly written and nothing can be understood.
Disfluent	The sentence is badly written and it takes a lot of effort to understand it.
Good	The sentence is partially understandable but has some small errors.
Flawless	The sentence has no errors and its of the sentence is very good.

2. You have to read the MT proposal and evaluate its adequacy. What does that mean? You must assign one of the following scores:

None	The meaning of the translation proposal contains no fragments that respect the text in the source language.
Little	The meaning of the translation proposal contains some fragments that respect the text in the source language.
Most	The meaning of the translation proposal almost completely respects the text in the source language.
Everything	The meaning of the translation proposal completely respects the text in the source language. There are no omissions or additions.

Annex 3. Guidelines for the Productivity evaluation

Objective

The aim of this document is to clearly define quality expectations and post-editing recommendations to evaluate an MT.

IMPORTANT!




In this task, you will post-edit a series of machine translated segments. The aim is to calculate how much time it takes to make the post-editing (therefore, you will be timed). We recommend that you post-edit as if it was a professional assignment, and that you post-edit all the segments in a go, as the milliseconds and number of editions will be counted. This way, we will be able to compare the the post-editing effort of different MT systems.

You can pause the task, but it is not recommended. This may influence the time calculation of your post-editing task. In addition, when you move from a segment, you won't be able to go back, so what you've done will be saved and it won't be possible to change.

The post-editing interface will be the one you see in the next screenshot, where you can see the current segment in the source text and the target segment with the MT raw output (in Catalan).

Information
Required Level of Quality: [Similar or equal to human translation](#)
Content Type: User Interface Text
Filename: PE_Sample1_TANS_taus_xlsx_empty_prod-qual.xlsx
Segment: 1 of 100

Source: English (United Kingdom)
Start
Current This entity does not have a unique ID, therefore its settings cannot be managed from the UI.
Next The {platform} integration is not loaded.

Target: Catalan
Start
Current Aquesta entitat no té un ID únic, per tant la seva configuració no es pot gestionar des de la IU. 
 
Or Press Enter

General Guidelines

- Raw MT output should not be completely deleted.
- Use as much raw MT output as possible.
- Check that the post-edited segment corresponds and faithfully reproduces the original text.
- No important information should be omitted or added.
- Check if the terms that have not been translated must remain untranslated.
- The grammatical and syntactic rules of the target language must be taken into account.
- Remember (if possible) the most recurring errors to comment them on later. This way, you will help improve the MT engine.

Practical Guidelines

Tone of voice (formal/informal):

When the user addresses the computer, the second person singular imperative form ("tu"; informal) will always be used. We will often find this in the menus, in some dialog boxes and on the buttons.

Original	Edit
Incorrect	Editar
Correct	Edita

When the computer/software addresses the user to provide information, ask a question, etc., the form to be used is the imperative in the second person plural ("vós"; formal). This tone of voice is to be used in documentation and in some dialog boxes.

Original	Choose the language you want to show this website in.
Incorrect	Tria la llengua en què vols mostrar aquest lloc web.
Correct	Trieu la llengua en què voleu mostrar aquest lloc web.

Locale information (numbers, measurements, currencies, dates, etc.):

Localization problems that may be caused because the MT engine failed to adapt the translation to the corresponding locale. The post-edited text should follow the rules of the target language.

Original	More than 250,000.75 people died because of this pandemic.
Incorrect	Més de 250,000.75 persones van morir a causa d'aquesta pandèmia.
Correct	Més de 250.000,75 persones van morir a causa d'aquesta pandèmia.

Localization problems:

The MT engine did not respect some of the key elements in the localization process, such as variables, links or non-translatable elements within a string.

Original	Hello {name}, welcome.
Incorrect	Hola, {nom}, benvingut.
Correct	Hola, {name}, et donem la benvinguda.

Accuracy problems and incorrect translations:

The target language text does not accurately represent the original text. The words or terms used are incorrect or too literal, or a proper name has not been respected.

Original	Home Assistant information
Incorrect	Informació de l'assistent d'estar per casa
Correct	Informació de Home Assistant

Punctuation errors

MT often tends to imitate the punctuation of the original text. Punctuation rules in the target language must be carefully checked.

Original	Nobody appears to have asked this “behavioural insights team.”
Incorrect	Ningú sembla haver preguntat a aquest “equip de coneixement de comportament.”
Correct	Sembla que ningú ha preguntat a aquest «equip de coneixement del comportament».

Use of uppercase letters

English tends to use more uppercase letters than Catalan. The natural use of capital letters in Catalan must therefore be followed, regardless of what the MT system suggests.

Original	When Prime Minister Boris Johnson held his first major press conference.
Incorrect	El Primer Ministre Boris Johnson va celebrar la seva primera gran conferència de premsa.
Correct	El primer ministre Boris Johnson va celebrar la seva primera gran conferència de premsa.