MT Evaluation in Translation Quotations at CPSL with ContentQuo

Lucía Guerrero
Machine Translation Specialist at CPSL
lguerrero@cpsl.com

Kirill Soloviev
CEO at ContentQuo
kirill.soloviev@contentquo.com

Abstract

Evaluating the performance of an MT system with new content - that is, MT performance prediction - is one of the most challenging aspects of MT, mainly because lack of reference translations does not allow using automatic metrics. Additionally, human evaluation can be expensive and time-consuming. As an LSP, at CPSL we deal with hundreds of translation requests daily and must choose the most appropriate workflow for our customers in a timely manner. That's why we needed a fast, reliable and cost-effective solution allowing us to find out if a given MT system is suitable for specific content. After trying different methods and tools, we chose the solution provided by ContentQuo, a translation quality management platform, based on the widely accepted Adequacy-Fluency methodology for MT evaluation. In our presentation we will introduce you to the challenges of MT quality evaluation and how we address them with ContentQuo.

1 Introduction

With over 50 years providing multilingual solutions to customers from around the globe, CPSL receives hundreds of translation requests on a daily basis. They may vary not only in terms of language combination but also of volume, deadlines, subject matter, expected quality levels and many more aspects. Quotation requests can be received and addressed both by dedicated Key Account Managers (KAMs) and Project Managers (PMs). The evaluation of the files received (DTP time, word count and other aspects to be considered) is done at the Technical and Evaluations Department (DTE) and shared with the KAM and the PM, who then produce a quotation.

Sending the quotation within the day is part of our standard practice. The aim is not only to stay competitive but also to provide our customers with accurate and timely information allowing them to take decisions.

In some cases, the KAM or the PM can consider machine translation as an option, either because the customer's budget is limited, because the deadline is very tight, because the user only needs the translation for assimilation (for understanding) or for several other reasons. Scope, time and cost are indeed considered the main drivers in the adoption of machine translation —they are also part of the "iron triangle" in project management (Muzii, 2016).

In such situations PMs and KAMs are encouraged to request what we call an "MT test" to the Machine Translation Specialist (MTS). The aim of this test is to determine if the raw MT of a given system is suitable as is or for a full or light post-editing (PE) of the file or files to be quoted; thus, if we can offer MT or MTPE to a customer in a given quotation.

In order to do this, the MTS must first assess a priori if the contents to be translated seem suitable for MT. For instance, texts which require transcreation are generally not considered for MT, whereas user manuals and in general texts written in a plain style, with a consistent use of terminology and without spelling or grammar errors are good candidates.

In the described scenario, there is no time or enough specific corpora to train and test a customized MT system. Therefore, based on the contents to be translated, the MTS must check which of the MT systems available could produce acceptable results, extract a sample and pre-translate it with

said system —actually in some MT tests the sample is pre-translated with more than one MT system for comparison.

2 Specific challenges

The difficulties which we encounter when evaluating MT in such circumstances are multiple. To start with, we cannot use automated metrics such as edit distance or BLEU. They compare the raw MT output with a reference translation. However, in cases such as the ones described, when a particular content type has never been translated before, there is no reference translation to compare with.

Second, and most importantly, we face time constraints. Normally the results are required in a very short timeframe, i.e. in the next 2-3 hours, because PMs and KAMs need to analyse the results and produce and send the quotation to the customer within the day. Finding the proper evaluator can be especially challenging for language combinations which cannot be covered by our inhouse Linguistic Leads, so sometimes it takes more time than expected to find someone immediately available for a 1-2 hours evaluation[1]. Cost constraints are also limiting of the ways in which this assessment is done. In the quotation phase, a Language Services Provider (LSP) does not know for certain if they would win the project or if their quotation will be rejected for any reason, so costs should be kept at a minimum. According to several industry-standard practices, such as the TAUS recommendations on evaluation and MT post-editing, samples should contain at least 200 segments and be evaluated by more than one person. However, these recommendations are not affordable in a real scenario. Finding a balance between cost and reliable results is key: we decided that our samples should be around 1500 words and unfortunately there is only time and budget for one evaluation.

Human evaluation, therefore, emerges as the only feasible option, but it has its own flaws, mainly the fact that it is subjective. In an academic setting this issue would be sorted out using several human evaluators. Unfortunately, as mentioned, time and cost limitations do not allow for this in our scenario. Furthermore, some evaluators report difficulties when annotating errors or evaluating MT, as well as lack of the proper tools and guidance to do such a job.

Finally, since samples need to be representative enough to allow for a meaningful evaluation, sampling can become an arduous task, especially when there are lots of files or source files come in a non-editable format.

3 Evaluation strategies

We have already mentioned that, for our purpose, we need to rely on the human evaluation of a representative sample of the contents. While it seems obvious that the sample should be sent in bilingual format (source and target need to be checked in order to be able to detect all types of errors, especially mistranslations), there are many different ways and tools for collecting the evaluator's feedback. Below we describe the strategies that were considered or adopted at CPSL before we tried ContentQuo, with their qualities and flaws according to our particular needs.

Quality Estimation (MTQE) is a feature offered by Memsource, one of the CAT tools we most use at CPSL, which calculates percentage quality scores automatically right after pre-translation and before any post-editing is done. It improves post-editing efficiency because it allows post-editors to focus on the segments showing worst quality scores. It is a feature that has proven very useful when the MTS already knows which MT engine should be used and the objective is to

1 It is outside the scope of this paper to discuss about the profile of the evaluators but it's worth mentioning that at CPSL we send MT tests to our most experienced post-editors, who are people used to work with machine translation and, therefore, do not have prejudices towards this technology

improve productivity, but not for the purpose of "MT tests": most part of the times there are hardly segments in the high quality categories and the rest fall into the "no score" category (i.e. post-editing is definitely required), which doesn't contribute to deciding about the suitability of a given system for a particular contents. It is anyway a feature to be further explored in the future.

Productivity tests are useful if there is a value to compare with. However, in "MT tests" most of the times we are required to evaluate an MT system for a type of content that has never been translated before. This means that the post-editing speed by itself will not tell us anything about how suitable a system is. It is also one of the less cost- and time-effective solutions.

Error annotation is a strategy in which the evaluator must report errors (if not all, then at least the most significative and frequent ones) and classify them according to a list of (usually) pre-defined categories: spelling, mistranslation, punctuation, grammar, style, etc.[2] Whereas it certainly offers valuable fine-grained information which the MTS can rely on to improve a given MT system, it is of no use at this point, in which we only need to know if a system is suitable or not. It is very time-consuming, and the resulting report does not allow for grasping an overall view of feasibility. At CPSL we began using error annotation in MT tests; however, we soon realized that this was not the correct approach: a list of mistakes alone is not enough to draw any conclusions about whether the output is suitable or not for post-editing.

Overall error scoring (also known as **document-level holistic evaluation**) is similar to error annotation but, in this case, the evaluator is simply required to assign an overall score to each category from a worst-best scale, based on the whole sample. Whereas it takes less time than the error annotation and the resulting report is easy to understand, the evaluators reported to the MTS that it was very difficult for them to come up with an overall score for each category by simply reading the pre-translated sample; thus, they tended to choose neutral midpoints which were of no use for the MTS.

It has already been mentioned that the raw MT output is always sent in a bilingual file; either an .xliff file processed in a CAT tool, or in a bilingual table, to allow the evaluator to look for mistranslations. The format in which the MT evaluators are required to compile their feedback is also relevant. They must receive the sample and the evaluation grid in a way which is clear and easy to understand and use. As commented, there is no time to learn how to use complex applications or copying and pasting from one file to another.

[2] Our list of categories is based on the MQM and TAUS DQF recommendations. Automatic error classification tools such as HJerson have not been considered so far.

CLIENT (name - key)	
PROJECT/QUOTATION	
LANGUAGE COMBINATION	
DOMAIN	
ANALYZED BY	
DATE	

Preliminary feedback report

LINGUISTIC ASPECTS		SCORE	Example (source)	Example (raw MT)	Example (file name)	Comments
Punctuation						
Spelling	Is it correct? Besides, if target is EN, indicate in Comments if spelling is US or UK					
Grammar	Concordance, verbs conjugation, etc.					
Terminology	Adequacy to context/precision?					
Completeness	Any omissions/additions?					
Semantics	Does raw MT convey the same meaning as the source text?					
Syntax	Word order, fluency, etc.					
Style						
COMPARISON TO HUMAN TRANSLATION						
	How close is the raw MT output to a high quality human translation?					
SCORE (1-5)	1 = Worst					
	2					
	3					
	4					
	5 = Best					

Figure 1: Template sent to evaluators to provide feedback on MT quality

When we moved from error annotation to overall error scoring, we designed the error categories and scoring grid in the same format file as the one we use for error annotation, a spreadsheet which we called 'Preliminary MT feedback report'. It was self-explanatory and easy to fill in and understand by the MTS, although it required copy-pasting from the sample to add examples.

To prevent file hassle, at a later stage we changed the format from a spreadsheet to an online form using Google Forms, which was shared with evaluators as a link in an email. The whole structure was the same but more questions (such as 'Is the raw MT output suitable for post-editing?') and fields for comments were added. The MT evaluators reported that they preferred online forms to spreadsheets because they allowed them to complete the task faster.

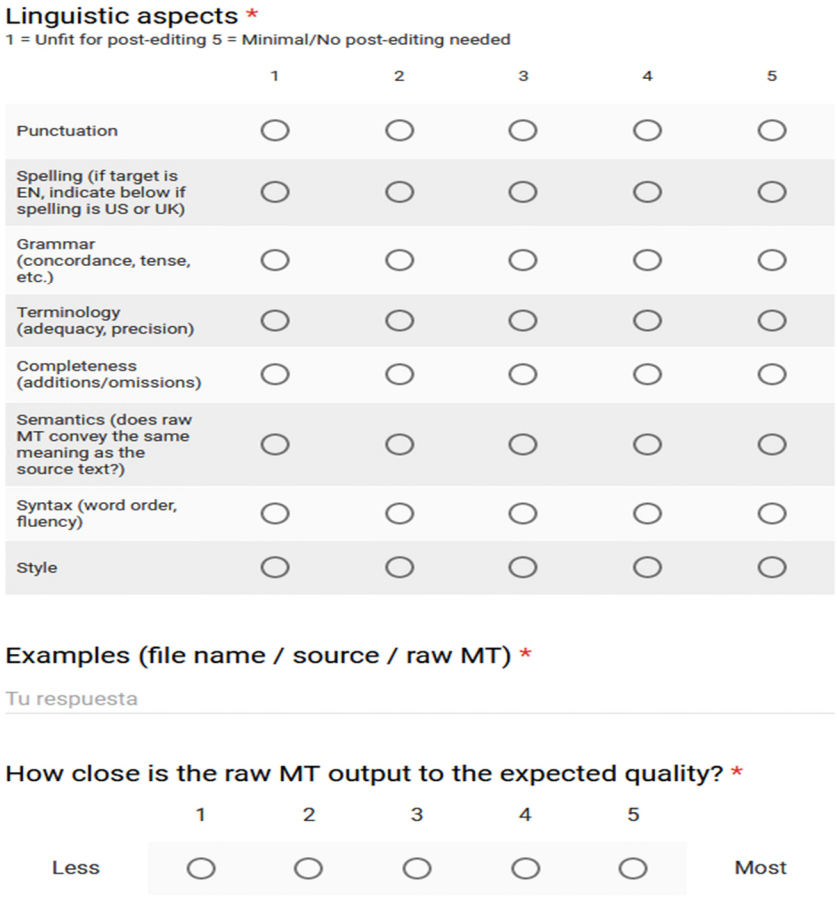


Figure 2: Online form sent to evaluators to provide feedback on MT quality

However, as soon as we started rolling MT massively into our company and MT tests were required almost daily, we realized that these online forms were falling behind with our needs. Google Forms did not allow for filtering the evaluations or obtaining any kind of meaningful statistics. Therefore, it was impossible to draw general conclusions about which system or systems were best at what type of contents, or if a given system had already been tried with a content type. The results of the evaluations were kept online and stored as a PDF in a folder, and almost forgotten after that.

It was the right moment to move to a centralized system which would allow us to filter all this valuable information, to **monitor all our MT tests and to offer an agile platform to our evaluators.**

4 Moving to ContentQuo

ContentQuo is a Translation Quality Management platform used by Global Top-10 Language Service Providers as well as commercial and government translation departments to reduce their linguistic quality risk, improve vendor performance, and boost the quality of Machine Translation output at scale with a data-driven approach. Their solution makes it easy to define, measure, analyse, and improve linguistic quality for both HT and MT alike —whatever the required quality measurement approach and/or translation technology stack is being used.

At CPSL we already had a ContentQuo licence which was being successfully used by the Production team (i.e. PMs and Linguistic Leads) to run regular quality checks on human translation in particular language combinations or content types, especially those with high quality requirements. It was being used to assess not only the quality of the texts per se but also the performance of translators and reviewers. It soon proved to be an excellent tool to track quality over time.

ContentQuo allows for choosing amongst an array of content profiles such as MQM or TAUS DQF, and even design your own. When they introduced support for segment rating quality measurement approach (such as Adequacy-Fluency) specifically designed for MT evaluation, we decided to be part of the first companies to try it. After testing it for two months, we decided to move to ContentQuo for all MT tests.

These are the features that we most value at CPSL:

4.1 Holistic Adequacy-Fluency profile

This quality profile is specifically designed to measure the quality when there is no time for error annotation. It combines a holistic profile with the Adequacy-Fluency approach recommended by TAUS, which consists of a human linguist evaluating each segment according to these two criteria.

Adequacy corresponds to "How much of the meaning expressed in the gold-standard translation or the source is also expressed in the target translation?". Evaluators are requested to rate, on a 4-point scale, how much of the meaning is represented in the translation:

- Everything
- Most
- Little
- None

Whereas **fluency** refers to: "To what extent is a target side translation grammatically well informed, without spelling errors and experienced as using natural/intuitive language by a native speaker?" In this case, evaluators must rate on a 4-point scale the extent to which the 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:

- Flawless
- Good
- Dis-fluent
- Incomprehensible"

This profile comes with these built-in values, but they can be customized.

The setup only takes a few seconds. The PM, LL or MTS, who have previously created ino the platform the profile of the MT systems and evaluators, must simply upload the bilingual sample, already pre-translated with the MT system chosen, to the ContentQuo, select the quality profile (Holistic Adequacy-Fluency in this case), the language combination, the kind of segments to be taken into account (e.g. ignoring full translation memory matches) and assign the translator (in our case, the MT system) and the reviewer (the evaluator). ContentQuo can even select a random sample from the file or files automatically if needed, which is very convenient when the quotation involves multiple files or long documents.

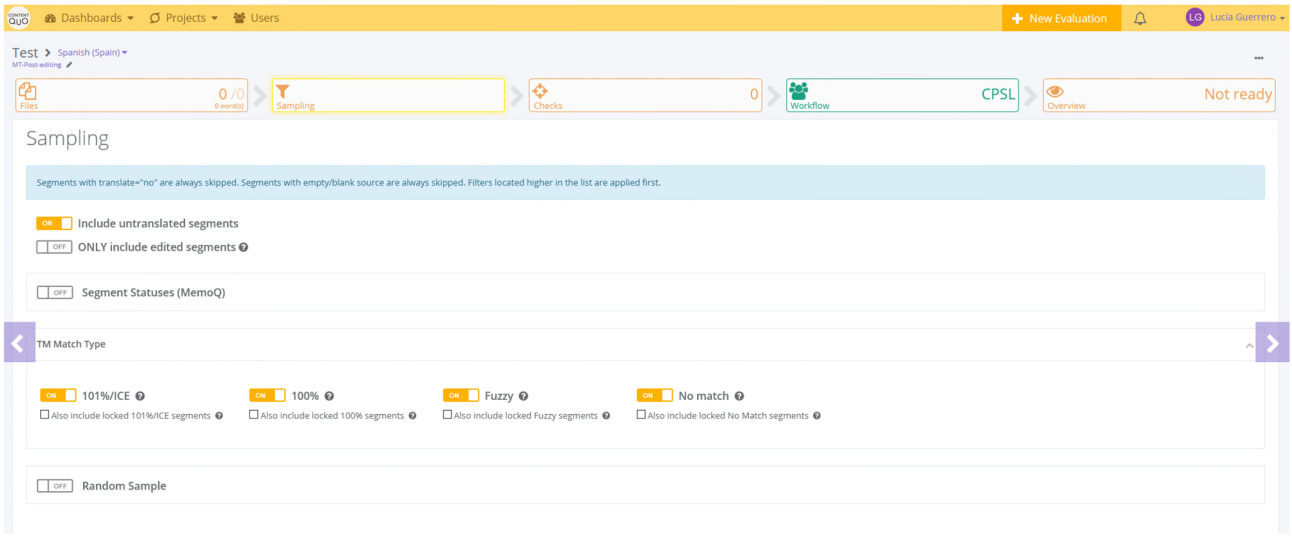


Figure 3: Defining the sampling and TM match settings of an evaluation in ContentQuo.

ContentQuo can be integrated with 3rd party automatic QA tools (like Xbench or QA Distiller) which allow for checking terminology consistency, spelling and many more potential issues. Such automatic checks can be of help to the MT evaluators when scoring segments.

Once everything has been setup, the requester can add specific instructions for the task and, by clicking the Start button, the evaluation is triggered, and an automated email is sent to the evaluator. The requester doesn't have to send anything else, other than the purchase order for the time invested in the evaluation.

4.2 Interface like a CAT tool

Everything is carried out online. Evaluators can read the MT output in their browser (source and target segment by segment), score segments one by one and (optionally) correct the errors, which will be conveniently displayed as tracked changes with no further action required.

Scoring is as simple as clicking the corresponding bullet. To understand the meaning of bullets, as well as the aspects covered by adequacy and fluency respectively, a pop-up message appears when placing the mouse over them. Finally, there is a field in which evaluators can add their comments.

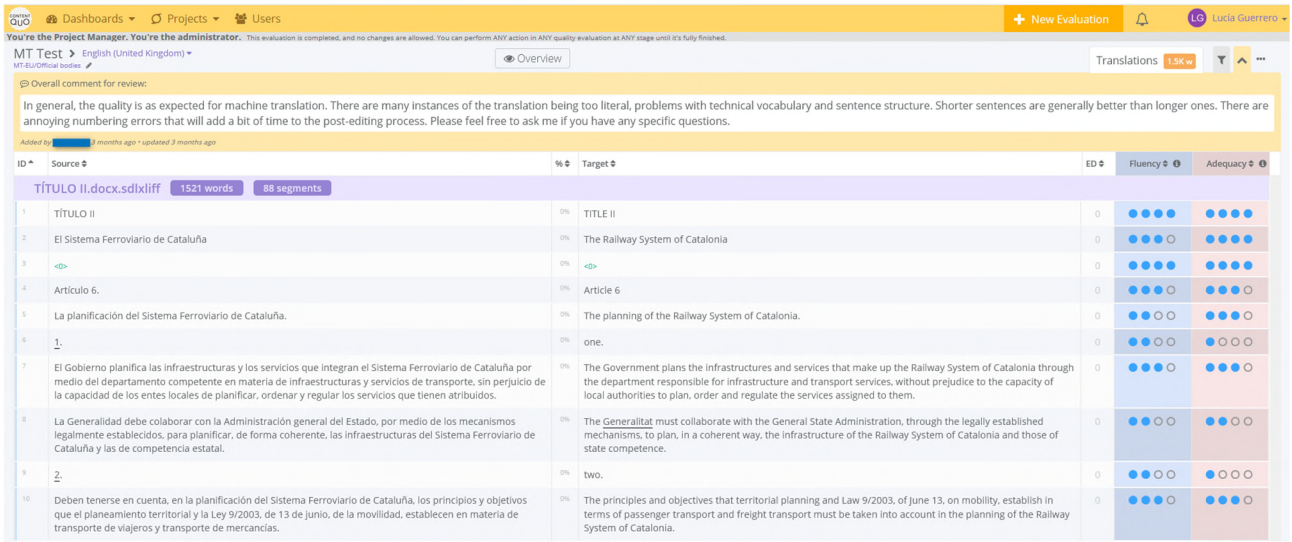


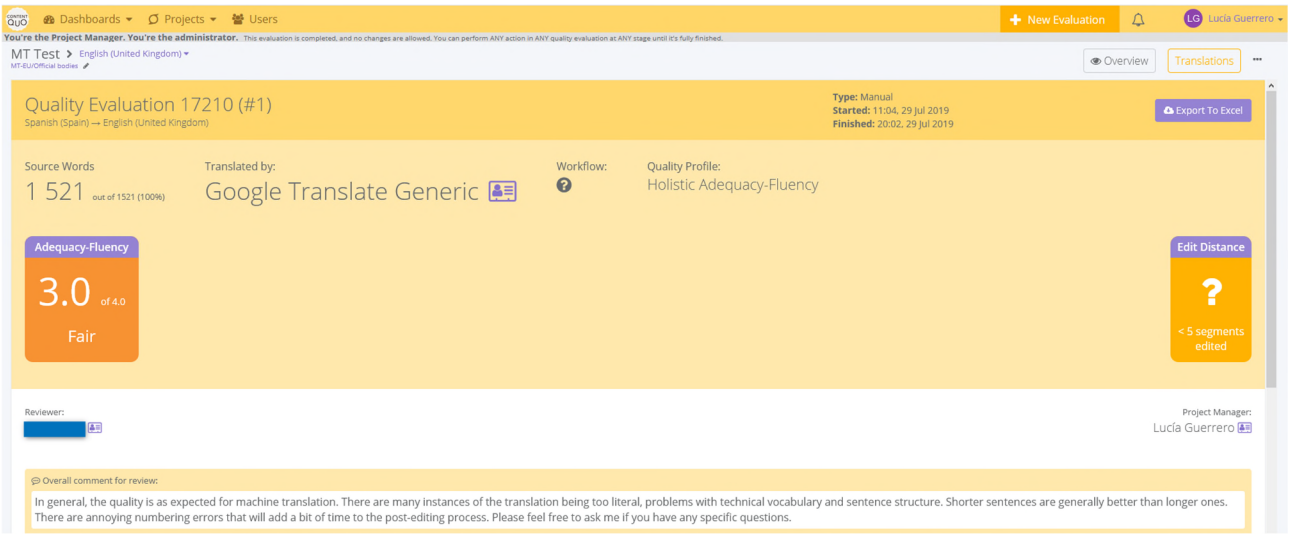
Figure 4: Segment by segment evaluation in ContentQuo.

4.3 Presentation of results

The MTS receives an email whenever the evaluation task has been completed with a link to the online results, based on the scores given to all segments, and they are displayed in a scorecard which can also be downloaded in spreadsheet format. The scoring thresholds we use at CPSL are as follows (as there is no industry-standard so far), and these can be customized as needed:

- 1.0-2.0 = Poor
- 2.0-3.0 = Fair
- 3.0-3.5 = Satisfactory
- 3.5-4.0 = Good

The scorecard in Excel format is shared with the post-editors if the quotation is approved, because it contains important information about the performance of the MT system. It is also accessible for them online.



Figures 5 and 6: scorecard with the evaluation results.

4.4 Analytics

Finally, all MT test results can be checked online at any time and filtered by language combination, MT system, evaluator and group (we are currently using the "Group" field to indicate the customer's name).

5 Conclusion

Post-editing of machine translation is an option which is worth exploring in several situations. Machine translation evaluation without a reference is only possible with human evaluation and needs to be done as quickly as possible, and in a cost-effective manner. Therefore, it is paramount to provide the evaluators with a system with which they can easily review the sample in bilingual format and work with a meaningful scoring system —all while reducing the manual effort and increasing the speed of delivering translation quotations that directly drives the company's business.

The feedback from our evaluators about using ContentQuo for MT tests was excellent from the beginning. They highlighted the ease of use, the user-friendly interface and most importantly the fact that, somehow, being able to score segments one by one made them feel like their evaluations were more reliable. As a member of our team puts it, "it's like adding quantitative data to qualitative feedback".

ContentQuo is so far the platform that best adapts to our needs at CPSL when it comes to MT tests. The MTS gets a quick overview of the suitability of a given MT system and, if needed, can still request a fine-grained evaluation on the same platform. Evaluators feel confident when submitting their scores, and consequently PMs and KAMs can trust the results and offer workflows tailored to the customers' needs.

As there is no perfect solution, there are currently a few improvements which we would like to see implemented soon, such as adding a time-control feature allowing for productivity tests, integration with more TMS and a specific field for subject matter (which would allow us to filter evaluations by this aspect as well). The last two features are already being considered and will be implemented in the upcoming months.

References

Burchardt, A. and Lommel, A. 2014. *Practical Guidelines for the Use of MQM in Scientific Research on Translation Quality*. Available at <http://www.qt21.eu/downloads/MQM-usage-guidelines.pdf>.

Farrús, M.; Cossta-Jussà, M.; Mariño, B. and Fonollosa, J. 2010. "Linguistic-based evaluation criteria to identify statistical machine translation errors", at *Proceedings of the 14th Annual Conference of the European Association of Machine Translation*. Available at <https://upcommons.upc.edu/bitstream/handle/2117/7492/Farr%C3%BAs2010.pdf>.

Muzii, L. 2016. *Post-editing of machine translation for translation project managers*. TAUS, De Rijp (The Netherlands).

TAUS Academy: "Best practices: Evaluate and Post-Editing". Available at <https://www.taus.net/academy/best-practices#evaluate>.

TAUS (Massardo, I.; Van der Meer, J.; O'Brien, S.; Hollowood, F.; Aranberri, N. and Drescher, K.). 2015. *MT Post-Editing Guidelines*. Available at <https://www.taus.net/think-tank/reports/postedit-reports/taus-post-editing-guidelines>.