Quality Evaluation Today: the Dynamic Quality Framework

A. Görög
TAUS

ABSTRACT

Translation quality is one of the key topics in the translation industry today. In 2011, TAUS developed the Dynamic Quality Framework (DQF) in an attempt to standardize translation quality evaluation. In this paper, we will describe common approaches to translation quality and introduce the TAUS framework for QE. We will show that the development of this framework, initiated by the industry, was necessary to fill the gap between theory and practice. In We will give a short summary of the survey on quality evaluation and DQF that was conducted in the summer of 2014 among users of the DQF tools. Finally, we will suggest some ways academia and industry could and should collaborate with each other in the field of quality evaluation in the future.

1. Introduction

Translation quality is one of the key concepts in the translation industry. Measuring and tracking translation quality is essential for all players of the industry: more and more translation vendors offer different types and levels of quality resulting in dynamic pricing; translation buyers are seeking to know whether their customized Machine Translation (MT) engine is improving and would like to compare different MT providers; finally, translators need to set the threshold of TM/MT matches at the most optimal levels. And these are just a few examples where translation quality becomes central and increasingly tuned to user satisfaction.

This said, translations are evaluated using one arbitrary model (usually error-typology) while ignoring the fact that several models are available for this purpose. In 2011, TAUS developed the Dynamic Quality Framework (DQF) in an attempt to standardize translation quality evaluation. Quality in DQF is considered dynamic since today's translation quality requirements change depending on content type, purpose and audience. DQF contains a rich knowledge base, resources on quality evaluation and a number of tools to profile and evaluate translated content. The framework is based on the assumption that the evaluation type selected should always match the content type, purpose, and communicative context of the given translation in a flexible, dynamic way. There is no one-size-fits-all approach to translation quality evaluation (QE).

In this paper, we will describe common approaches to translation quality (section 2) and introduce the TAUS framework for QE (DQF). We will show that the development of this

framework, initiated by the industry, was necessary to fill the gap between theory and practice (section 3). In section 4, we will give a short summary of the survey on quality evaluation and DQF that was conducted in the summer of 2014. Finally, we will suggest some ways academia and industry could and should collaborate with each other in the field of quality evaluation in the future (section 5).

2. What is translation quality?

Quality is when the user or customer is satisfied. A longer and more scientific definition of quality is as follows:

"A quality translation demonstrates required accuracy and fluency for the audience and purpose and complies with all other specifications negotiated between the requester and provider, taking into account end-user needs." (Melby, 2014, forthcoming)

Unfortunately, quality measurement in the translation industry is still not always linked to customer satisfaction and specifications. Very often, quality evaluation is the task of quality managers on the supply and demand side who have one specific evaluation model. This model is often based on error-typologies that assign different weights to different error types. Input from customers is usually missing or ignored and every translation receives the same treatment.

Despite very detailed and strict error-based evaluation models, it seems that satisfaction levels with both translation quality and the evaluation process itself are low. The major problem is that models and metrics used are not always measuring the right thing. Little consideration is given to multiple variables such as content type, communicative function, end user requirements, context, perishability, or mode of translation generation (whether the translation is created by a qualified human translator, unqualified volunteer, machine translation system or a combination of these). Traditional one-size-fits-all approaches to quality do not satisfy buyers and vendors of translation services anymore. QE models such as the LISA (Localisation Industry Standards Association) QE model, the J2450 or the EN15038 do not seem to take into account the different varying user requirements, communicative goals and content types. Are existing (ISO, LISA, ASTM, etc.) standards and certificates to ensure quality then useless. Standards that certify the translation vendor and/or the quality management process that lie beneath are certainly useful but cannot give a 100% guarantee that the product itself meets the required quality level. The only way to ensure that is by evaluating and doing that exactly the same way each time preferably across the whole industry. Today, there is an increasing appetite for such an approach to quality within the industry, an approach that measures the right quality level with the right method.

To offer such an approach and to standardize human evaluation of translated content, TAUS created the Dynamic Quality Framework (DQF). The DQF platform consists of a rich knowledge base on Quality Evaluation with best practices, reports, templates and a number of tools to evaluate translations made both by human translators and MT engines. The tools enable evaluators to compare translations, assess their accuracy and fluency, to measure post-editing productivity and to score translated segments based on an error typology. The Content profiling wizard enables users to select best-fit evaluation methods.

3. The Dynamic Quality Framework

3.1. Aim

Quality in DQF is considered dynamic as translation quality requirements change depending on the content type, the purpose of the content and its audience. The Framework provides a commonly agreed approach to select the most appropriate translation quality evaluation model(s) and metrics depending on specific quality requirements. The underlying process, technology and resources affect the choice of the quality evaluation model.

The Framework is underpinned by the recognition that quality is when the customer is satisfied. It is used when creating or refining a quality assurance program. DQF provides shared language, guidance on process and standardized metrics to help users execute quality programs more consistently and effectively. Improving efficiency within organizations and through supply chains. The result is increased customer satisfaction and a more credible quality assurance function in the translation industry.

3.2. Development

The development of DQF started in January 2011 by over fifty companies and organizations. Contributors include translation buyers and vendors as well as academic institutions. Users continue to define requirements and best practices as they participate in regular (online) meetings and events. Since the end of 2014, DQF is part of the TAUS Evaluate platform.

In the first phase of DQF development, TAUS carried out a benchmarking exercise to review evaluation models and this showed that existing QE models are relatively rigid[1]. For the majority, the error categories, penalties applied, pass/fail thresholds etc. are the same no matter what communication parameters were involved. The models are also of such a detailed nature that applying them is time-consuming and evaluation can only be done for a small sample of words. No standard tool was used for sampling neither for quality evaluation at the time. What's more, QE models are predicated on a static and serial model of translation production, which doesn't match 21st century expectations of dynamic pricing.

DQF offers a more flexible approach to the common static quality evaluation models since it is based on the three parameters of utility, time and sentiment (UTS). This model considers the communication channel – Regulatory, Internal, or External (B2C, B2B, C2C). It is informed by the results from the content profiling exercise performed by TAUS enterprise members collaborating in this project, which shows that it is possible to map content profiles to the evaluation parameters utility, time and sentiment.

[1] https://www.taus.net/reports/translation-quality-evaluation-is-catching-up-with-the-times

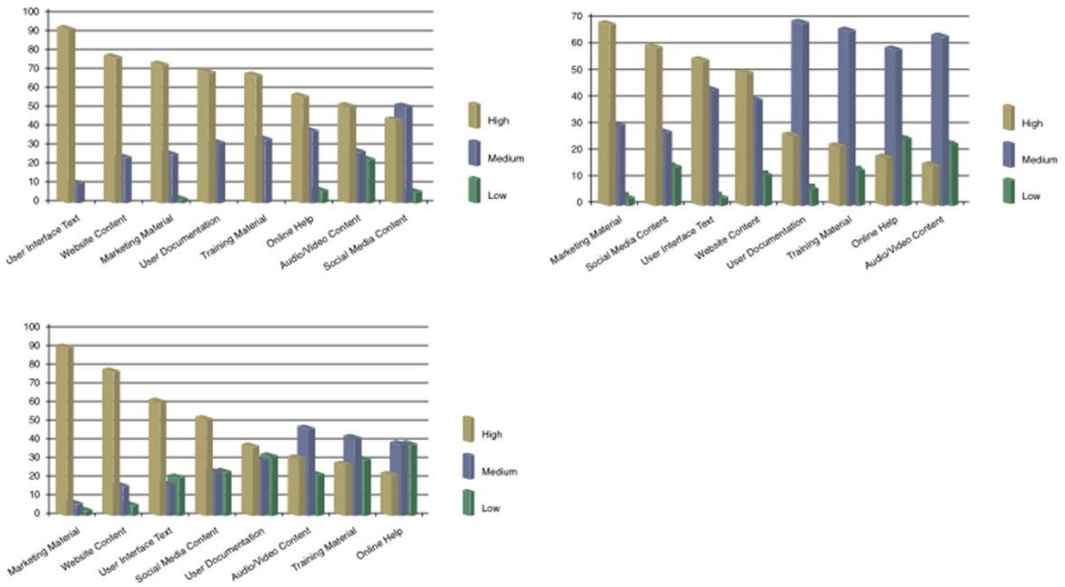


Figure 1: Importance of Utility, Time and Sentiment attributes distributed according to content types

The results of the content-profiling exercise also suggest that there are clear content differentiators for utility and sentiment while the parameter of time is much fuzzier. Reason is that most companies require a quick turnaround time for translations. Some examples of the mapping between content types and UTS rating are as follows:

- User Interface text and website content are rated highest for utility while audio/video content is rated lowest
- Marketing material and social media content are rated highest for time while user documentation, training material, online help and audio/video content are rated of medium importance for time.
- Marketing material and website content are given highest importance for sentiment while training material and online help are rated lowest for this parameter.

The Content Profiling wizard available on the TAUS Evaluate platform is one of the results of the TAUS benchmarking exercise described above. The DQF Content Profiling feature is used to help select the most appropriate quality evaluation model for specific requirements. This leads to the Knowledge base where you find best practices, metrics, step-by-step guides, reference templates, and use cases.

3.4. DQF tools

The DQF tools provide a vendor independent environment for the human evaluation of translation quality. Users gather vital data to help establish return-on-investment, measure productivity enhancements, and benchmark performance, helping to ensure that informed decisions are made. One of the aims of DQF tools is to standardize the evaluation process and make it more objective and transparent. The benchmarking and reporting functions provide users with a wealth of information on quality problems related to certain language pairs, text types, industries or domains.

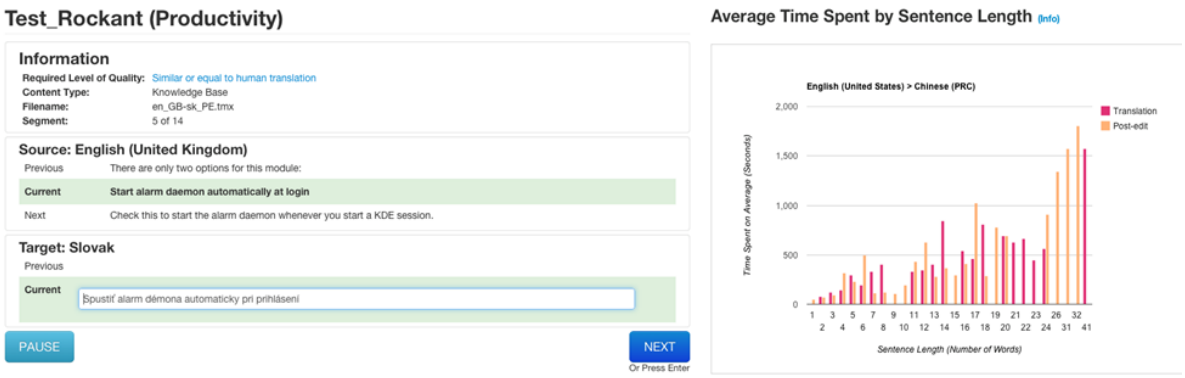


Figure 2: Productivity testing and reporting in DQF

DQF tools are created with the non-technical user in mind. The interface is extremely user friendly, which also makes it into an excellent teaching aid (more on this in *section 5*). The project manager creates a project, defines the evaluation task and uploads the translation file(s). The evaluators receive an email and begin the task. When the task is completed, the project manager receives an email asking to review the results. After clicking through, automatically generated reports are provided. Data can also be downloaded to create customized reports. The project manager can discuss the findings with the evaluators or compare the results to previous findings.

3.4.1. MT ranking and comparison

The Comparison Task helps users select MT engines or human translators based on the quality of the output. DQF limits the number of sources you can compare to three. Shared experience at TAUS member companies has shown that an evaluator's ability to make robust judgments is impaired if he or she has to score more than 3 options segment-by-segment. After the translation files are uploaded, evaluators are invited to compare the translated segments and to give a ranking. The tool randomizes the order in which the target segments are presented. This means the evaluator(s) do not get conditioned into giving anticipated rankings.

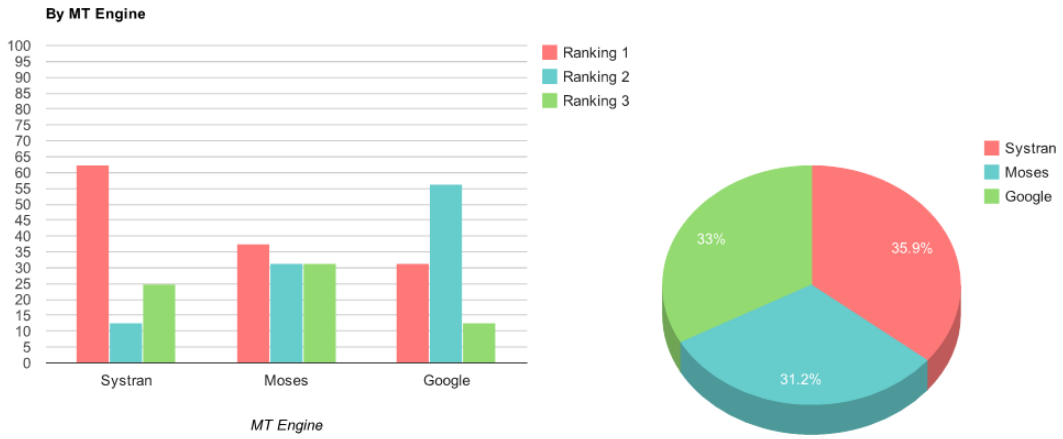
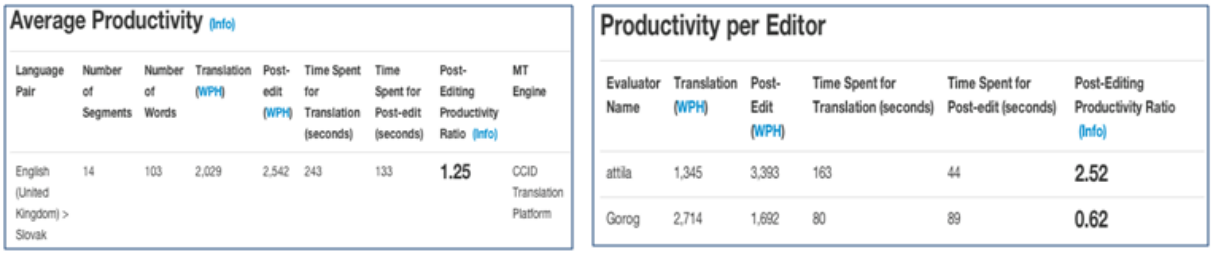


Figure 3: DQF MT ranking results

At the end of the task, the project manager can see which engine or translator yields better results for a certain language combination on a given text-type. Users can also gain insight into common errors.

3.4.2. Productivity testing

Post-editing productivity testing is becoming one of the most practical ways of generating evaluation scores. This evaluation type enables you to assess the difference in speed between MT post-editing and translating from scratch. This DQF tool removes half the target side (MT output) segments from your uploaded file(s). Users therefore have to translate half the segments from scratch and edit the other half. The system measures the time taken to complete these tasks. When assigning the task to users, you need to specify which of the two types of post-editing is required (i.e. light or full).



Average Productivity (Info)

Language Pair	Number of Segments	Number of Words	Translation (WPH)	Post-edit (WPH)	Time Spent for Translation (seconds)	Time Spent for Post-edit (seconds)	Post-Editing Productivity Ratio (Info)	MT Engine
English (United Kingdom) > Slovak	14	103	2,029	2,542	243	133	1.25	CCID Translation Platform

Productivity per Editor

Evaluator Name	Translation (WPH)	Post-Edit (WPH)	Time Spent for Translation (seconds)	Time Spent for Post-edit (seconds)	Post-Editing Productivity Ratio (Info)
attila	1,345	3,393	163	44	2.52
Gorog	2,714	1,692	80	89	0.62

Figure 4: Productivity results in the DQF reporting tool

The results provide insight into the difference in time and effort between light and full post-editing. Users will also learn about the impact of certain errors on translation quality, the variance across languages and content types, the correlation with certain metrics and scores or the influence of the translator's profile (age, gender, experience, etc.) on post-editing. This test can also be used to compare different MT engines in a more indirect way. DQF also offers the possibility to post-edit the whole text offering time-measurements and edit-distance information at the end of the evaluation project making DQF a simple, user-friendly post-editing environment with productivity reporting.

3.4.3. Error typology

Error typology is the standard approach to quality evaluation currently. There is some consistency in its application across the industry, but there is also variability in categories, granularity, penalties and so on. The DQF error typology tool offers a standardized way to categorize and count translation errors using commonly used industry criteria for accuracy, language, terminology, style and country standards. The DQF error typology was developed by considering existing error-count metrics (such as the LISA QA Model). Another example of an error typology is the Multidimensional Quality Metrics (MQM) developed in the European QT Launchpad project and owned by DFKI.

Tracking and comparing the errors found in computer-generated translations offers insights into the weaknesses of MT engines and MT in general. Besides, a comparison of SMT with RBMT based on an error typology can be an interesting exercise that makes the differences between the two types engines more tangible for users. TAUS has published best practice guidelines on the error-typology approach. These guidelines enable users to adopt standard approaches to error typology evaluation, ensuring a shared language and understanding between translation buyers, suppliers and evaluators.

As of September 2014, TAUS and DFKI have started the harmonization of DQF and MQM with the aim of bridging the gap between the definitions and specifications of the two models. TAUS acts as the industry outreach platform for the harmonized model for all stakeholders: translators, language service providers, translation buyers, government institutions and NGOs. TAUS will continue offering access to the harmonized models through direct contacts with its membership as well as through partnerships with other associations and members. Both TAUS and DFKI has decided to reach out to and work with standards organizations like ISO and ASTM to share the harmonized model and offer the agreed specifications in the standardization process.

3.4.4. Adequacy/ Fluency

This evaluation type is in use in machine translation evaluation and can be equally adopted for human translation quality evaluation. It involves measuring two text quality attributes:

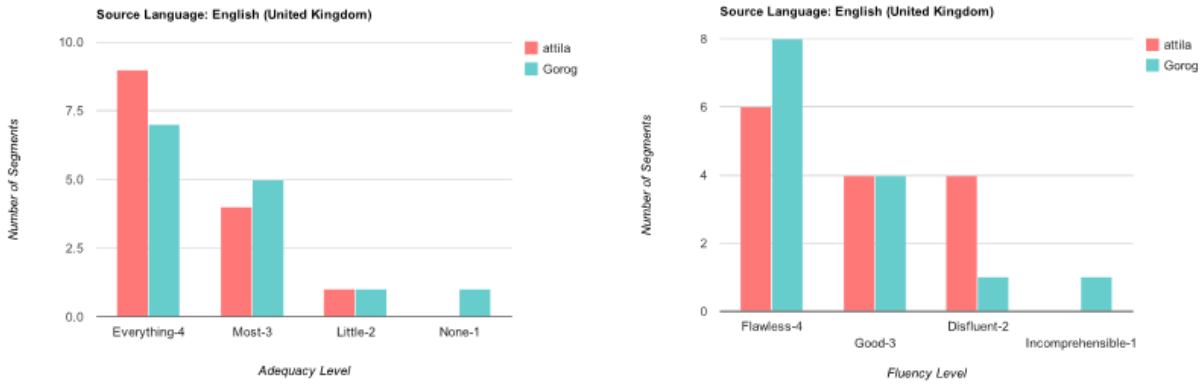


Figure 5: Adequacy and fluency diagrams showing the results per evaluator

In DQF, the definition of the Linguistic Data Consortium is used for Adequacy and Fluency.

- Adequacy: "How much of the meaning expressed in the gold-standard translation or the source is also expressed in the target translation."
- Fluency: To what extent the translation is "one that 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"

4. Quality evaluation today

In the summer of 2014, a survey was carried out by TAUS. The aim was to collect user feedback on TAUS DQF and on Translation QE in general. By that time, more than 500 regular and irregular users had been using the DQF tools to evaluate translation quality. The full survey will be published for public consultation on the TAUS website by mid-December 2014.

A third of the participants in this survey were translators themselves or academic staff. They were using DQF as an evaluation tool to evaluate MT output or human translations. In-house staff (the largest percentage of respondents – 38%), freelance translators or linguists using the DQF tools, on the other hand, were doing that commissioned by LSPs or translation service buyers. They were conducting MT comparison/ranking to help decide which MT solution is best

fit for the purpose. A smaller percentage (5%) of the users at the buyer side outsourced translation QE to another company, usually to another LSP or a specialized linguistic consultancy firm while another 14% chose to outsource evaluation to freelance translators.

Respondents noted that evaluators of translation quality do not always have the necessary skills. Very often untrained translators, interns or bilingual staff members are asked for the job. Training and providing ample information prior to an evaluation project are essential. A video might serve as a first step to provide information and training. The downside of this is that a video might also be difficult to consult as a quick reference on an ongoing basis. Generally, some written instructions are needed. They should be brief and positive (Do's rather than Don'ts). Additional instructions could be provided depending on the translation evaluation purpose. The advantage of written pointers is that they can easily be consulted at any stage. Providing examples can be also very useful.

On the methodology side, the difficulty of differentiating between preferences and errors, the lack of consistency, objectivity and time were pointed out as major problems. Subjectivity is a major concern when it comes to evaluation even when the TAUS DQF tools are used. While DQF offers a transparent and standardized workflow for translation QE, increasing the number of evaluators as well as providing the necessary instructions beforehand, can improve the credibility of the evaluation results. The decision on how detailed evaluation should be is necessary before every evaluation process. One can choose for a monolingual fluency evaluation or a bilingual accuracy evaluation or a combination of these two topped with an error-typology evaluation. Different types of evaluations and different combinations require different amount of time and expertise to conduct.

Finally, participants would like to see a definition emerging for good quality and for different quality levels. They mentioned the lack of transparent evaluation criteria when conducting evaluations in existing tools and metrics. Finding the right metrics already causes problems to some. Others complained about the lack of standardized sampling algorithms. Though human evaluation remains a valuable method for assessing translation quality, it can be time-consuming and expensive. A recent model to obtain quicker and cheaper human evaluations is by means of sampling. Sampling can be used in several situations within the translation workflow; the two main use cases pointed out during the breakout session at the TAUS QE Summit Dublin 2014[2] being Quality Assessment of human translation and machine translation evaluation. Sampling is appropriate in the most evaluation scenarios, but the scenario in which sampling takes place has a strong influence on how the sample will be designed and analyzed. Sampling can also be a useful technique in translator training and continuing development.

It was interesting to see the variety of tools and methods applied by the respondents before starting to use TAUS DQF including automated metrics (WER, BLEU), Excel forms, open source tools (e.g. Appraise), the LISA QA model and in-house tools. TAUS's approach to QE was justified by the recommendation of some respondents that urged the industry to use an agreed set of standards and metrics (such as DQF) since quality is something every Language Service Provider (LSP) offers to clients but without defining or measuring it in a proper way.

[2] https://www.taus.net/taus-quality-evaluation-summit-2014

5. DQF in training and research

Although translation QE has always been an essential part of the translation process in the industry, it is only now that it's gaining importance in academic research. Translation quality evaluation data enables researchers to answer several of the following questions: what exactly are the key features of good content and how can we measure them? What are the general problems in enabling machines to 'understand' language? Which text types are most amenable to MT? What are the advantages and disadvantages of different MT approaches (RbMT vs SMT)? How can we compare two translations of the same source text in a consistent way? How can a user improve the performance of an MT system? What are the requirements of effective post-editing?

Since DQF is freely available for academic research and education purposes, an increasing number of universities have been using the tools. DQF tools and reports enable researchers to investigate the achievements and limitations of (commercially available) MT systems such as Google translate, Bing etc. They can also assess which text types are suitable for processing with these technologies. And they can also evaluate human translations or compare post-editing to translation from scratch. Although DQF is free for research, large volumes of evaluation data are still missing. Work in the area has been hampered by the lack of availability of relevant data to train metrics. Companies are not keen on offering their data to research purposes even though this type of data is often abundant among providers and buyers of automatic translations, since they routinely need to assess translations for quality assurance. Research on better automatic evaluation metrics would therefore greatly benefit from a closer relationship between industry and academia.

Platforms and tools such as TAUS DQF (Dynamic Quality Framework) can facilitate such collaboration between industry and academia by providing systematic ways of collecting and storing quality assessments (according to specific requirements for a given content type, audience, purpose, etc.) that can be directly used to train metrics. Additionally, quality evaluation and quality estimation could be integrated into such platforms to support human evaluation. Academia needs to obtain more feedback, information and requirements from the industry to better focus research activities on solutions to the problems that the industry is actually facing. Industry also needs better software solutions from academia, both in terms of usability and performance, in order to test the techniques and solutions designed by the industry.

6. Conclusion

Since TAUS launched the Dynamic Quality Framework in 2011, we have learned to apply different methods of QE such as adequacy, fluency, productivity testing and MT ranking. We have also learnt to compare results to previous projects and to minimize subjectivity by using a standardized workflow. What's still missing is benchmarking to satisfy user needs and to provide the right level of quality for each user. In order to develop and improve translation quality, we need to measure quality constantly and consistently. But how can we achieve that when budgets and resources set-aside for this purpose are so tight. How to become efficient in QE? Using DQF tools, users can now research the achievements and limitations of their MT engines. They can assess which text types are most suitable for processing with these technologies. They can also

evaluate human translations or compare post-editing to translation from scratch. The final aim, of course, remains satisfying the customer.

References

Alan Melby: 2012 LACUS lecture (2014, forthcoming)

Lena Marg, Sharon O'Brien, Attila Görög, Miguel Gonzalez: TAUS Best Practices on Community Evaluation

Luigi Muzii: Quality Assessment and Economic Sustainability of Translation

Sharon O'Brien, Rahzeb Choudhury, Jaap van der Meer, Nora Aranberri Monasterio: TAUS Dynamic Quality Evaluation Framework: TAUS Labs report

Sharon O'Brien: Towards a Dynamic Quality Evaluation Model for Translation

Sharon O'Brien: Translation Quality - It's time that we agree