revista.tradumàtica
tecnologies de la traducció

443-454

Quantifying and benchmarking quality: the TAUS Dynamic Quality Framework

Attila Görög
TAUS

ABSTRACT

Translation quality is one of the key topics in the translation industry today. In 2011, the Translation Automation User Society (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. DQF is freely available for academics and can be accessed on the TAUS Evaluate platform.

Keywords: translation quality evaluation, MT evaluation, DQF, TAUS, tools, metrics, post-editing productivity

RESUM *(Quantificació i avaluació comparativa de la qualitat: el Dynamic Quality Framework de TAUS)*

La qualitat en traducció és un dels temes clau actualment a la indústria de la traducció. El 2011 la Translation Automation User Society (TAUS) va desenvolupar el Dynamic Quality Framework (DQF) en un intent de normalitzar l'avaluació de la qualitat en traducció. DQF concep la qualitat de manera dinàmica, ja que actualment els requisits de qualitat en traducció canvien en funció del tipus de contingut, la seva intenció i el seu destinatari. DQF conté una àmplia base de coneixement, recursos sobre avaluació de qualitat i un gran nombre d'eines que permet perfilar i avaluar el contingut d'una traducció. DQF està disponible de manera gratuïta per a investigadors i s'accedeix a través de la plataforma d'avaluació de TAUS.

Paraules clau: avaluació de la qualitat en traducció, avaluació de TA, DQF, TAUS, mètriques, productivitat en postedició.

RESUMEN *(Cuantificación y evaluación comparativa de la calidad: el Dynamic Quality Framework de TAUS)*

La calidad en traducción es uno de los temas clave actualmente en la industria de la traducción. En 2011 la Translation Automation User Society (TAUS) desarrolló el Dynamic Quality Framework (DQF) en un intento de normalizar la evaluación de la calidad en traducción. DQF concibe la calidad de manera dinámica, ya que actualmente los requisitos de calidad en traducción cambian en función del tipo de contenido, su intención y su destinatario. DQF contiene una amplia base de conocimiento, recursos sobre evaluación de calidad y un gran número de herramientas que permite perfilar y evaluar el contenido de una traducción. DQF está disponible de manera gratuita para investigadores y se accede a través de la plataforma de evaluación de TAUS.

Palabras clave: evaluación de la calidad en traducción, evaluación de TA, DQF, TAUS, métricas, productividad en posedición.

1. Introduction

Translation is a complex linguistic process and quality is probably the most complex variable in this process. While specifying the cost and measuring the speed of translation are trivial, assessing the quality of translated documents is much more difficult. If a translation is evaluated today in the industry, it is often done using one arbitrary model ignoring the fact that several models are available. DQF[1] by TAUS[2] 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 way. There is no one-size-fits-all approach to translation quality evaluation (QE).

While 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. By taking on quality evaluation and research on translation quality on a greater scale, be it machine-generated or made by human translators, we are able 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 Machine Translation (MT)? How can we compare two translations of the same source text in a consistent way? What are the most appropriate and reliable techniques for evaluating translated content? What are the requirements of effective post-editing?

In this paper, we will describe common approaches to quality evaluation (**section 2**) and introduce the Dynamic Quality Framework showing how the development of this framework was initiated by the industry and how it is filling a gap (**section 3**). We will explain the different evaluation tools available in DQF (**section 4**) and we will discuss how benchmarking is becoming essential when it comes to comparing translation quality at an industry level (**section 5**). Finally, we will suggest some ways academia and industry could and should collaborate in the field of quality evaluation (**section 6**).

2. Quality Evaluation anno 2014

Today's changing views on translation quality is a hot topic for all players of the industry: translation buyers aim for different types of quality and flexible ways of pricing; Language Service Providers (LSP) are keen to know whether their customized MT engine is improving; and translators would like to set the threshold of Translation Memory (TM)/MT matches at the optimal levels. And these are just a few examples of where translation quality is becoming more attuned to different user needs.

Quality is when the user or customer is satisfied. A longer and more academic definition of quality is (Melby, 2015, forthcoming): "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". Unfortunately, quality measurement in the translation industry is still not always linked to customer satisfaction. Very often, quality evaluation is the task of quality managers on the supply side who have specific evaluation models. These models are often based on error typologies that assign different weights to different error types without any input from customers.

Quality evaluation is problematic. 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. According to O'Brien et al. (2011), one of the problems is that models and metrics used are not always measuring the right thing. Little consideration is given to

[1] http://www.evaluate.taus.net/evaluate/dqf/dynamic-quality-framework

[2] TAUS is a resource center for the global language and translation industries. Its mission is to increase the size and significance of the translation industry to help the world communicate better. http://www.taus.net

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).

In the industry, there has also been a recent focus on what constitutes acceptable levels of quality for different types of content and purposes. This new approach to quality replaces the centuries old assumption that translation users always need the highest quality. Concepts such as "fit for purpose" and "good enough" translation have been supported by leading figures in the industry (Drugan, 2013; Prioux & Rochard, 2007).

Traditional one-size-fits-all approaches do not satisfy buyers and vendors of translation services anymore. QE models such as the LISA QA model[3], the SAE J2450[4] or the BS EN15038[5] do not seem to take into account different content types, varying user requirements and communicative goals (O'Brien, 2012). Today, there is an increasing appetite for a new approach to quality within the industry, an approach that measures the right thing with the right method. As a result, translation QE needs to re-focus on a number of cost-effective, practical issues (Muzii, 2006).

3. TAUS Dynamic Quality Framework

3.1 Aim

To optimize 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.

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.

DQF Tools, Content Profiling, Resources and Knowledge base are 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. The result is increased customer satisfaction and a more credible quality assurance function in the translation industry.

[3] The LISA model was developed by the Localization Industry Standards Association (LISA). The model includes error categories, related subcategories as well as severity and penalty points. World Wide Web documentation available at http://dssresources.com/news/1558.php

[4] SAE J2450 Society of Automotive Engineers Task Force on Translation Quality Metric, 1999. World Wide Web documentation available at http://www.sae.org/standardsdev/j2450p1.htm

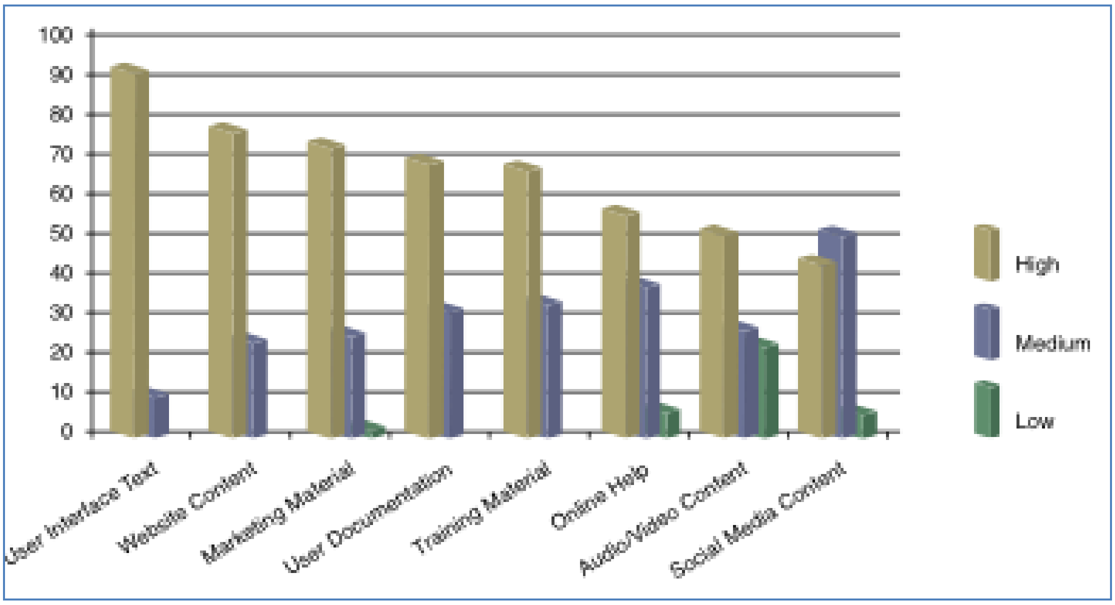
[5] BS EN-15038 is a European standard for translation services which covers the core translation process and all other related aspects involved in providing the service, including quality assurance and traceability. World Wide Web documentation available at http://qualitystandard.bs.en-15038.com

3.2 Development

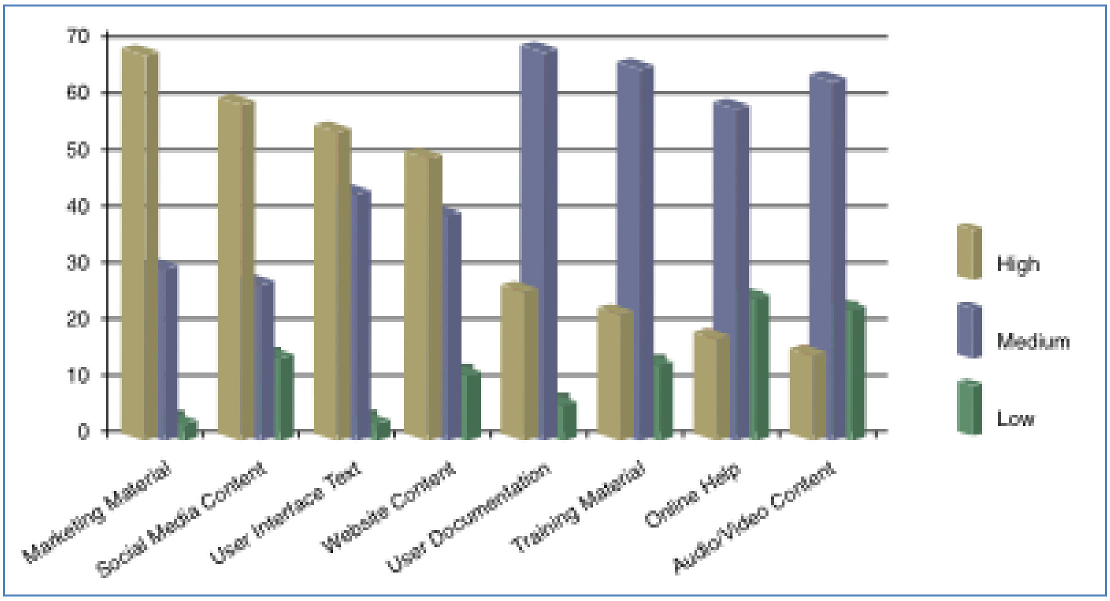
DQF has been co-created since January 2011 by over fifty companies and organizations. Contributors include translation buyers, translation service providers, translation technology suppliers and academic institutions. Practitioners continue to define requirements and best practices as they evolve through regular meetings and events. Since 2014, DQF is part of the online TAUS Evaluate platform[6].

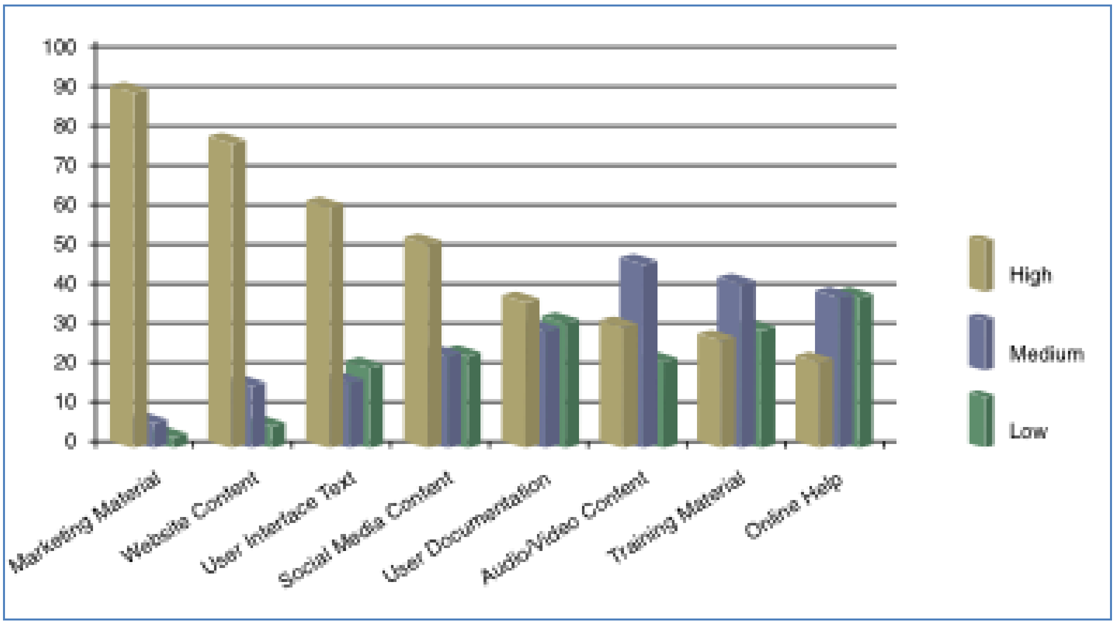
TAUS carried out a benchmarking exercise in Q1 2011 to review existing evaluation models (O'Brien et al., 2011) and this showed that existing QE models are relatively rigid. For the majority, the error categories, penalties applied, pass/fail thresholds, etc. are the same no matter what communication parameters are 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 or for quality evaluation at that time. Furthermore, existing QE models were predicated on a static and serial model of translation production, which is not suited to the emerging models of ubiquitous computing.

DQF offers a more flexible approach than these static quality evaluation models since it is based on the three parameters of Utility, Time and Sentiment (UTS). In Translation quality evaluation, Utility refers to the relative importance of the functionality of translated content, Time to the time with which the translation is required and Sentiment to the importance of impact on brand image, i.e. how potentially damaging might it be to a translation client if content is badly translated.



[6] https://evaluate.taus.net



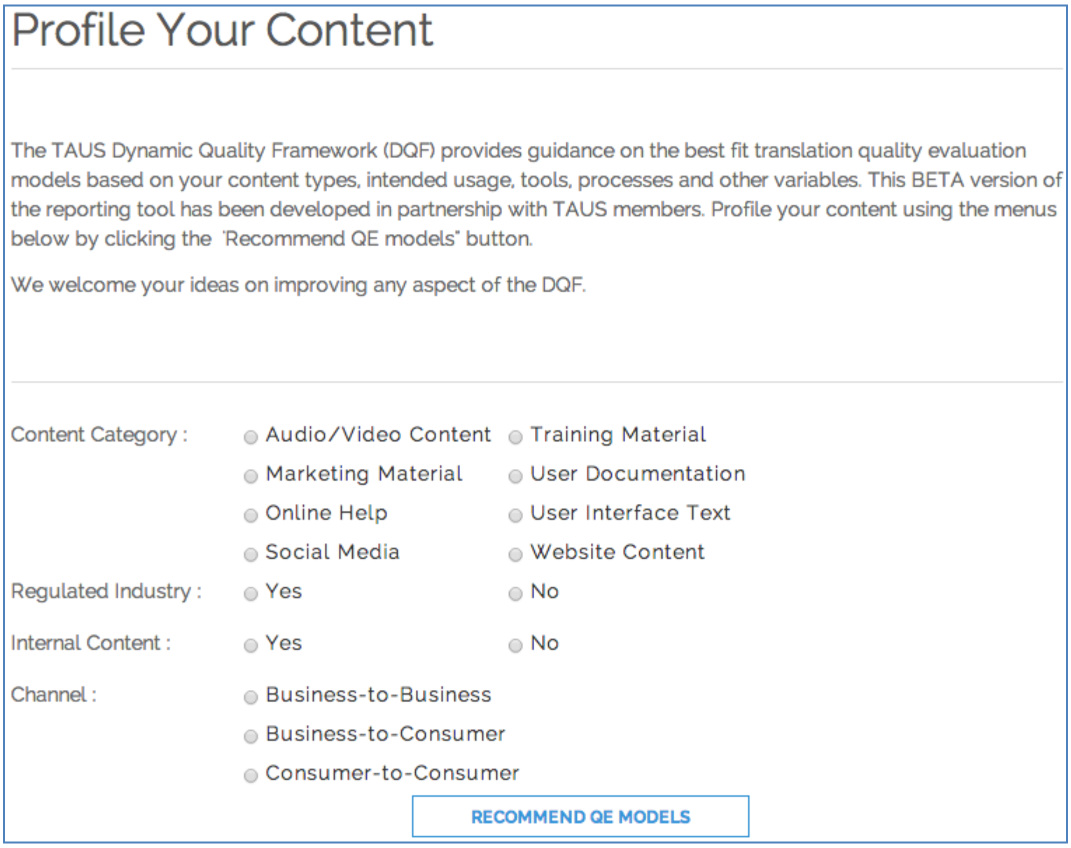


Figures 1-3. Importance of Utility, Time and Sentiment attributes distributed according to content types.

A dynamic QE model should also consider the communication channel. There is a wide variety of communication channels in use today. The communication channel might be internal to a company (e.g. internal training material) or external. For external communication, it is suggested that there are at least three channels for the flow of translated content:

- B2C: Business-to-Consumer
- B2B: Business-to-Business
- C2C: Consumer-to-Consumer.

The C2C model caters for multi-lingual user-generated content, which is consumed by other multi-lingual consumers (e.g. tweets, blogs, user forum postings etc.). A B2C communication channel will presumably require a stricter evaluation model than a B2B or C2C model. Quality expectations for internal communication might be lower than for external communication and so on. Ultimately, the quality model cannot be divorced from the communication channel (O'Brien, 2012).



Profile Your Content

The TAUS Dynamic Quality Framework (DQF) provides guidance on the best fit translation quality evaluation models based on your content types, intended usage, tools, processes and other variables. This BETA version of the reporting tool has been developed in partnership with TAUS members. Profile your content using the menus below by clicking the 'Recommend QE models' button.

We welcome your ideas on improving any aspect of the DQF.

Content Category : ○ Audio/Video Content ○ Training Material ○ Marketing Material ○ User Documentation ○ Online Help ○ User Interface Text ○ Social Media ○ Website Content

Regulated Industry : ○ Yes ○ No

Internal Content : ○ Yes ○ No

Channel : ○ Business-to-Business ○ Business-to-Consumer ○ Consumer-to-Consumer

Figure 4. TAUS DQF Content Profiling wizard including attributes for Content Category, Regulated Industry, Internal Content and Communication Channel

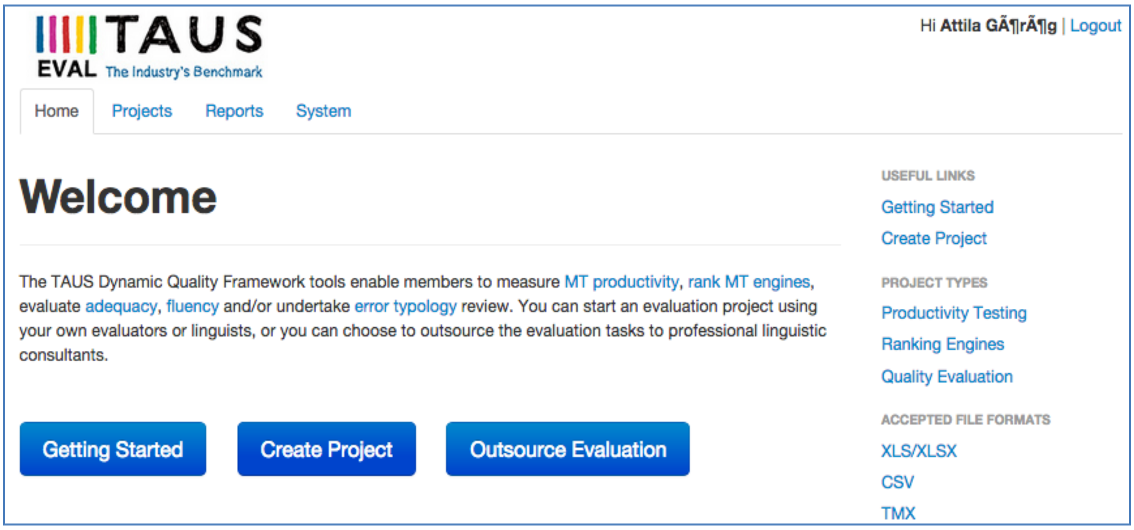
The results of the content-profiling exercise also suggest that there are clear content differentiators for utility and sentiment while the parameter of time is much fuzzier. The reason is that most companies require a quick turnaround time for translations. Some examples of the mapping between content 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 the highest importance for sentiment while training material and online help are rated lowest for this parameter.

Taking these results, DQF proposes that UTS ratings can be mapped on to specific QE models to recommend the most suitable model for each user's needs. This is the basis of the TAUS Content Profiling tool.

4. DQF tools

Once the user is ready with profiling the translated content and selected the most appropriate evaluation type, the evaluation project can be set up in the 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 to 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.

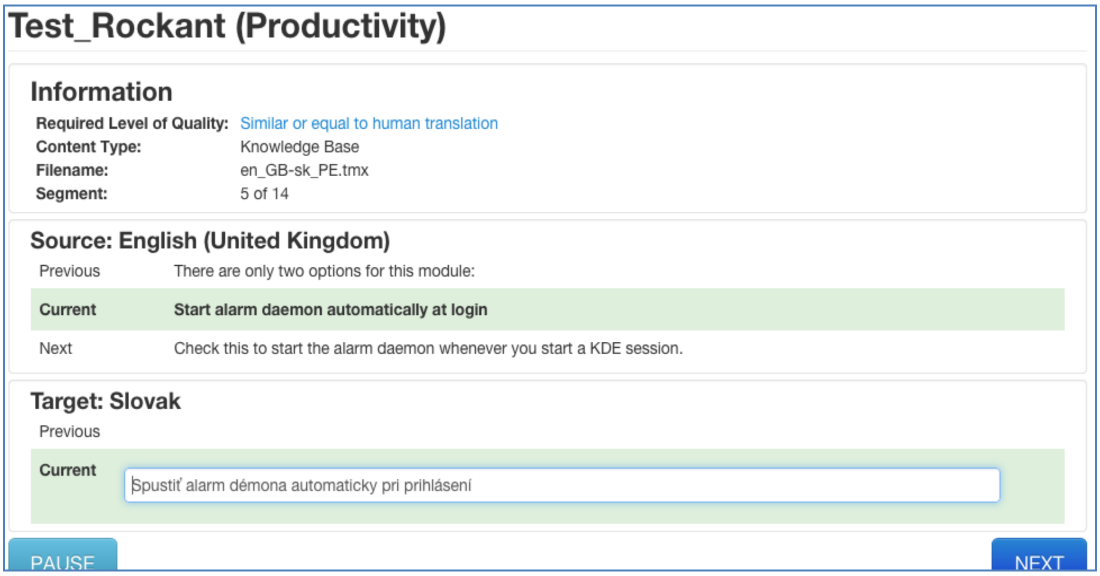


Welcome

The TAUS Dynamic Quality Framework tools enable members to measure MT productivity, rank MT engines, evaluate adequacy, fluency and/or undertake error typology review. You can start an evaluation project using your own evaluators or linguists, or you can choose to outsource the evaluation tasks to professional linguistic consultants.

Figure 5. Landing page of the TAUS DQF tools

DQF tools are created with the non-technical user in mind. The interface is extremely user friendly which makes it into an excellent teaching aid too. 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. The results can also be downloaded to create customized reports. The project manager can discuss the findings with the evaluators or compare them to previous findings.



Test_Rockant (Productivity)

Information

Required Level of Quality: Similar or equal to human translation
Content Type: Knowledge Base
Filename: en_GB-sk_PE.tmx
Segment: 5 of 14

Source: English (United Kingdom)

Previous	There are only two options for this module:
Current	**Start alarm daemon automatically at login**
Next	Check this to start the alarm daemon whenever you start a KDE session.

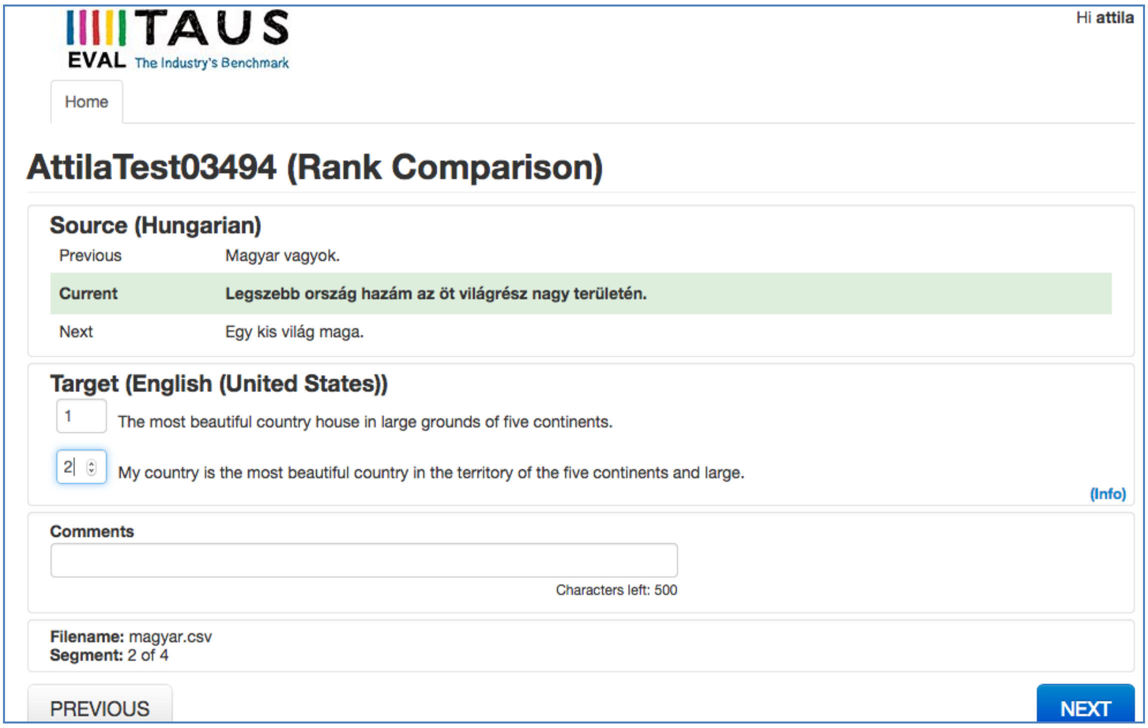
Target: Slovak

Previous	
Current	Spustiť alarm démona automaticky pri prihlásení

Figure 6. User-friendly interface in the TAUS DQF tools for productivity tasks

4.1 MT ranking and comparison

The Comparison Task helps users select MT engines based on the quality of the output. DQF limits the number of engines you can compare to three. Research 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 from the engines being compared are presented. This means the evaluator(s) do not get conditioned into giving anticipated rankings. At the end of the task, the project manager can see which engine yields better results for a certain language combination on a given text-type. Users can also gain insight into common errors made by MT engines. Evaluators can also be asked to compare two or three human translations of the same source text and rank the translation segment-by-segment.



AttilaTest03494 (Rank Comparison)

Source (Hungarian)

Previous Magyar vagyok.

Current **Legszebb ország hazám az öt világrész nagy területén.**

Next Egy kis világ maga.

Target (English (United States))

1 The most beautiful country house in large grounds of five continents.

2 My country is the most beautiful country in the territory of the five continents and large.

Comments

Filename: magyar.csv
Segment: 2 of 4

Figure 7. Rank comparison task in the TAUS DQF tools

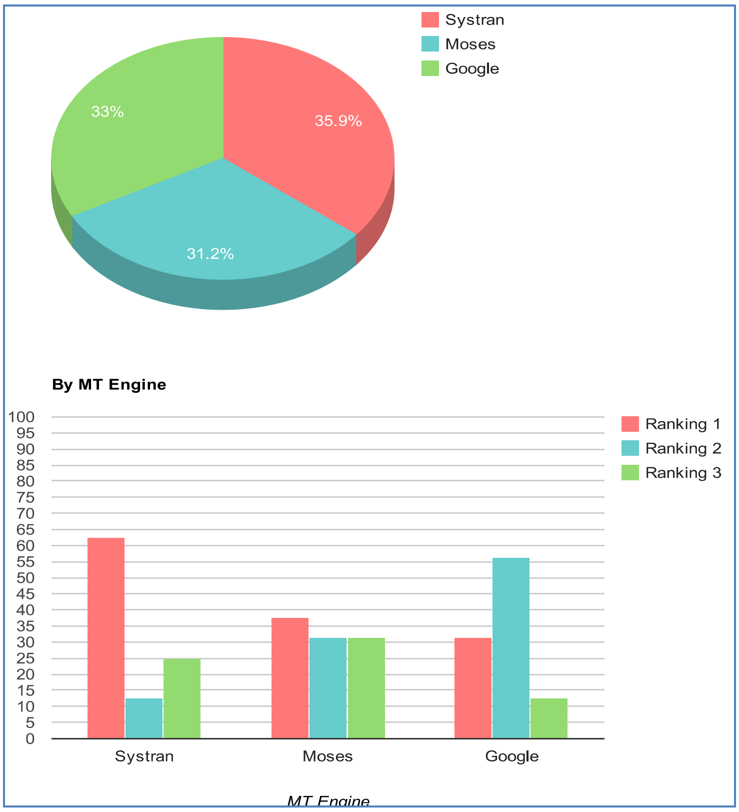


Figure 8. Results of an MT ranking evaluation project

4.2 Post-editing productivity testing

Post-editing productivity testing is becoming one of the most practical ways of generating evaluation scores. In this evaluation, users can choose to either post-edit the entire MT-output or translate half of the segments from scratch and post-edit the other half. In the latter case, the DQF tool removes half the target side (MT output) segments from the uploaded file(s). In both cases, the system measures the edit distance[7] and the time taken to complete the tasks. When assigning the task to users, you need to specify which of two types of post-editing is required (i.e. light or full)[8]. 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.

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. Considering existing error-count metrics, with the LISA QA Model playing a central role, TAUS developed the DQF error typology.[9]

Another example of an error typology can be found in the Multidimensional Quality Metrics (MQM) developed as part of the European QTLaunchPad project. MQM provides a framework for describing and defining quality metrics used to assess the quality of translated texts and to identify specific issues in those texts. It provides a systematic framework to describe quality metrics based on the identification of textual features. MQM is intended to provide a set of criteria to be used to assess the quality of translations.

Tracking and comparing errors found in computer-generated translations offers insights into the weaknesses of MT engines and MT in general. Furthermore, a comparison between Statistical Machine Translation (SMT) and Rule-Based Machine Translation (RBMT) based on an error typology can be an interesting exercise that makes the differences between the two types of engines more tangible for users. Error-typology based evaluations are also applied to human translation mostly as part of a review environment or with the purpose of error-annotation.

As of September 2014, TAUS and the Deutsche Forschungszentrum für Künstliche Intelligenz GmbH (DFKI)[10] have started the harmonization of DQF and MQM with the aim of bridging the gap between the definitions and specifications of the two models. Once the

[7] Levenshtein's algorithm is used to calculate edit distance. (See also: http://en.wikipedia.org/wiki/Levenshtein_distance)

[8] To reach quality similar to "high-quality human translation and revision" (a.k.a. "publishable quality"), full post-editing is usually recommended. For quality of a lower standard, often referred to as "good enough" or "fit for purpose", light post-editing is usually recommended. More information on light vs. full post-editing can be found in the TAUS/CNGL Machine Translation Post-editing Guidelines available on the TAUS website.

[9] The categories in LISA QA are: mistranslation, accuracy (omission, addition, cross-references), terminology (glossary adherence, context), language (grammar, semantics, punctuation, spelling), style (general style, register/tone, language variants), country (country standards, local suitability), and consistency. The penalty points in relation to the severity level are as the following: 1 = minor errors; 5 = major errors; 10 = critical errors.

[10] The German Research Center for Artificial Intelligence, with sites in Kaiserslautern, Saarbrücken, Bremen (with an associated branch in Osnabrück) and a project office in Berlin, is the leading German research institute in the field of innovative software technology. http://www.dfki.de

harmonization of the two models is completed, TAUS will act as an outreach platform for the harmonized model by offering access to all stakeholders of the translation industry.

5. Benchmarking

Creating objective and widely accepted benchmarks to compare translation quality on an industry level would make it possible to offer different levels of quality to buyers of translation services. This already happens sporadically. There are vendors today offering translation services "tailored to your needs": from budget translations through first draft translations to professional translations and transcreation. Based on the content and the purpose, you pay for the quality level you choose. Some LSPs ask customers to specify the quality level of the required translation. The different levels may vary from free to expert translation, with a definition of the different levels.

Collecting evaluation data on as much content, from as many evaluators and sources as possible, would give us an idea of the thresholds for different levels of quality. But how do you prove that your translation reaches a certain level? Is a quality mark for translations a viable option in today's translation landscape? We need to know the exact quality of the final product. Moreover, the customer needs to be able to specify the quality level he or she desires and can afford. Most customers have no clue what to say when they are asked to specify their desired quality level.

Offering different levels of quality only makes sense if there are commonly accepted quality levels or benchmarks and translations are evaluated against these benchmarks by independent, third party language services. Outsourcing quality evaluation is likely to become much more common in the translation industry in the near future.

6. DQF in research

Since the beginning of 2014, DQF has been freely available for the academia. The result is that an increasing number of academic institutions have been making use of DQF for both research and education. Using DQF tools, researchers can explore the achievements and limitations of (commercially available) MT systems. They can also assess which text types are most 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 purposes, 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 translations, since they routinely need to assess the quality of their translated content. 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 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 currently facing. The 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.

7. Conclusion

Since TAUS launched DQF in 2011, TAUS members have learnt to apply different methods of QE such as adequacy, fluency and productivity testing. They have also learnt to compare results to previous projects and to minimize subjectivity by using a standardized workflow. The new challenge is, however, to be able to compare evaluation results consistently across the whole industry. There is a need for benchmarking to satisfy user needs and to provide the right level of quality for each customer. The last word has not been said about this delicate topic, but one thing is certain: this problem cannot be solved in isolation.

In order to develop and improve translation quality, it needs to be measured constantly and consistently. But how can we achieve that when budgets and resources set aside for this purpose are so tight. How can we become efficient in QE? Unfortunately, there's no such thing as a free lunch and solving the QE bottleneck is one of the major challenges in our industry today. Assessing the quality of a translation can sometimes cost even more than producing the translation itself! Nonetheless, tracking delivered translation quality and sharing evaluation data are indispensible for automating the quality evaluation process. Industry and academia should start working together on achieving this aim. The quality of existing MT solutions can only be improved if we also implement better metrics and manage to automate (at least part of) the QE process. Without that, no advances will be made in the translation industry.

Bibliography

Drugan, J. (2013) Quality in Professional Translation, Bloomsbury, Leeds, UK

Melby, A. (2015, forthcoming) *2012 LACUS lecture*

Muzii, L. (2006) Quality Assessment and Economic Sustainability of Translation, In: *Rivista internazionale di tecnica della traduzione*, Issue 9, 15-38.

O'Brien, S., Choudhury, R., Van der Meer, J., Aranberri Monasterio, N. (2011) TAUS Dynamic Quality Evaluation Framework: TAUS Labs report, [online] https://www.taus.net/think-tank/reports/evaluate-reports/translation-quality-evaluation-is-catching-up-with-the-times

O'Brien, S. (2012) Towards a Dynamic Quality Evaluation Model for Translation, In: *Journal Of Specialised Translation*, Issue 17, 55-77

O'Brien, S. (2014) Translation Quality - It's time that we agree, [online] https://www.taus.net/think-tank/articles/event-articles/translation-quality-it-s-time-that-we-agree

Prioux, R. and Rochard, M. (2007) Economie de la révision dans une organisation internationale; le cas de l'OCDE in: *The Journal of Specialised Translation*, Issue 8, 21-41

TAUS/CNGL (2010) Machine Translation Post-editing Guidelines, [online] https://www.taus.net/think-tank/best-practices/postedit-best-practices/machine-translation-post-editing-guidelines