

The TAUS Quality Dashboard

Paola Valli

TAUS

paola@taus.net

Abstract

This workshop outlines the progress that has been made on the TAUS Dynamic Quality Framework (DQF) in the past year and introduces the TAUS Quality Dashboard where all stakeholders in the global translation services can monitor their performance using industry-shared metrics and benchmark themselves against industry average productivity and quality. The TAUS DQF integration with translation tools via an open API will also be demonstrated.

1. Introduction

The diversification in content types and the swift adoption of translation technologies (including machine translation) drives the need for more dynamic and reliable methods and measurements for translation quality evaluation. Industry-shared metrics will lead to more reliable measurements that give all stakeholders in the language service industry useful benchmarks and insights to help them adjust and improve their processes. The industry-shared metrics will turn quality evaluation into business intelligence steering and supporting management decisions.

In this workshop, we are going to present the progress that has been made on the TAUS Dynamic Quality Framework (DQF) since the last AsLing workshop one year ago. The workshop will also introduce the TAUS Quality Dashboard, which was released in September 2015. The Dashboard is an industry collaborative platform for the global translation services sector where translator operators and producers will be able to monitor their performance based on a variety of parameters they can select from.

We are going to present and demo the integration of DQF in CAT tools as well as the reporting features in the Quality Dashboard and are looking forward to receiving feedback and comments from the participants on the work already done and the future roadmap.

2. The Dynamic Quality Framework

The TAUS Dynamic Quality Framework (DQF) was first developed by TAUS in 2011 in close cooperation with many of the TAUS member companies and represents a dynamic approach to quality evaluation. This dynamic evaluation model takes into account the changing landscape accounting for different content types and the adoption of automated translation technologies. The theoretical framework of DQF is built around three evaluation parameters: utility, time and sentiment. The relative weight of these parameters varies in relation to the content type to be translated. The vision behind DQF is to standardize the methods and tools of quality evaluation, aggregate the scores and measurements and make these available through industry-shared metrics. While DQF provides the reference for quality evaluation, the DQF online platform, also known as DQF tools, provides the specific tools needed to carry out quality evaluation in a vendor independent and standardized environment. The DQF tools running on the TAUS website were released in 2014.

At the first AsLing conference last year, TAUS presented the results of a survey conducted in the summer of 2014 among translators and academic staff who were conducting quality evaluation tasks for MT output or human translation. All respondents were active

users of the TAUS DQF tools and were asked to provide feedback and explain how they did translation Quality Evaluation (QE) and what they expected from the DQF (van der Meer & Görög, 2015). Some of the points raised concerned the lack of transparent evaluation criteria, the difficulty of finding the right metrics, the lack of standardization and the need for different quality levels, not to mention costs and time-to-market.

The tools on the TAUS Evaluate platform include a Content Profiling wizard, a tool to carry out MT ranking and comparison, a tool to run post-editing productivity testing and a knowledge base containing best practices and use cases. Quality attributes for MT output are traditionally accuracy and fluency. However, accuracy and fluency can just as easily be adopted to evaluate human translation which can also be checked for types of errors, as the standard approach to quality evaluation currently does. DQF adopts the error typology developed from the existing error-count metrics (see Section 6).

3. From DQF to the Quality Dashboard

Collecting quality data through the DQF tools proved to be useful but at the same time this approach still suffered from the limitations of displaying only the data that were related to the submitted projects. If collected data could become shared metrics, measurements would become more reliable and give translation operators and producers (translators) useful benchmarks and insights that help them to adjust and improve processes. This is why a new perspective was taken as to what DQF could achieve.

TAUS members and partners started to ask whether there was a way of integrating DQF into the translation workflow and avoid the continuous switching between the normal environment and the DQF tools page. This is why an open API for DQF was developed that connects DQF to the existing translation tools and workflow systems. TAUS provides API specifications and dedicated plugins to allow technology providers and users of translation services to integrate TAUS DQF into their work environment.

The data collected through DQF can be displayed on the TAUS Quality Dashboard to allow translators and project, vendor and quality managers track and benchmark the quality, productivity and efficiency of translation.

The Quality Dashboard was a natural next step that fits very well with the overall trend in the industry towards open data and metrics. The Quality Dashboard delivers on the DQF vision and provides statistics on translation, benchmarking for translation activity and quality, as well as analysis of translation performance and production. Quality evaluation through the Quality Dashboard becomes business intelligence to help steer and support management decisions.

4. Reporting in the Quality Dashboard

The reports in the Quality Dashboard cover the two main areas of Productivity/Efficiency and Quality. [These two areas will be covered in more detail in the following sections]. The Quality Dashboard is a flexible and dynamic tool which offers a number of filters to customize the charts and reports to be displayed. At each level, users can see the overall industry average and the industry average for their specific selection. In addition, users can also benchmark their project(s) against the industry scores.

Available filters include language pair, time span, project, technology use (e.g. TM vs.

MT), translation process, content type and industry. Reports for quality will include error typology both in terms of number and type of errors. In addition, error review can be customized with penalties and pass/fail rates. There is a development roadmap for all reports to be made available to users and planned until the end of the year. Thanks to all the available filters, reports can be made more or less granular and additional filters can be developed on request from the users.

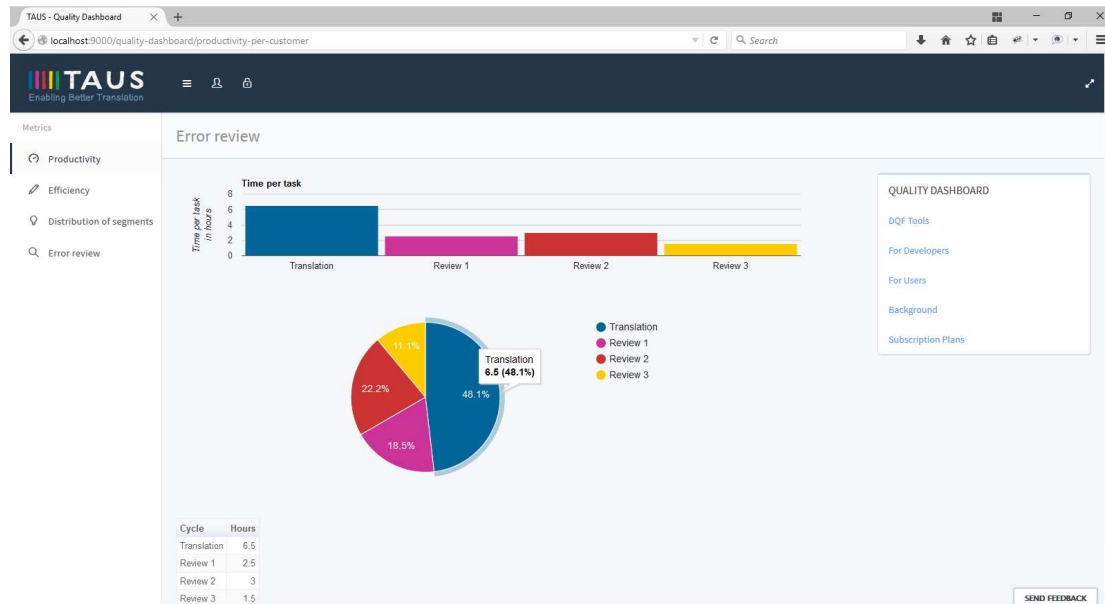


Figure 1 - Time spent per task

Figure 1 shows the productivity for each task of a project. In this case, productivity is expressed in total hours spent on translation and review, broken down by review cycle. Time spent per task can be displayed both in aggregated form per project or broken down e.g. per language pair. This allows the identification of possible bottlenecks in the overall workflow.

In addition to total number of errors per error category, another report can be generated which provides a more accurate picture of the distribution of errors based on their severity. Figure 2 shows how many errors per category have been labeled as ‘critical’, ‘major’, ‘minor’ or ‘neutral’ at project level, but the same information can also be provided for an individual task. The chart provides the weighted distribution (bars) compared to the absolute count (blue line). Both counts are normalized (e.g. per 1,000 words).

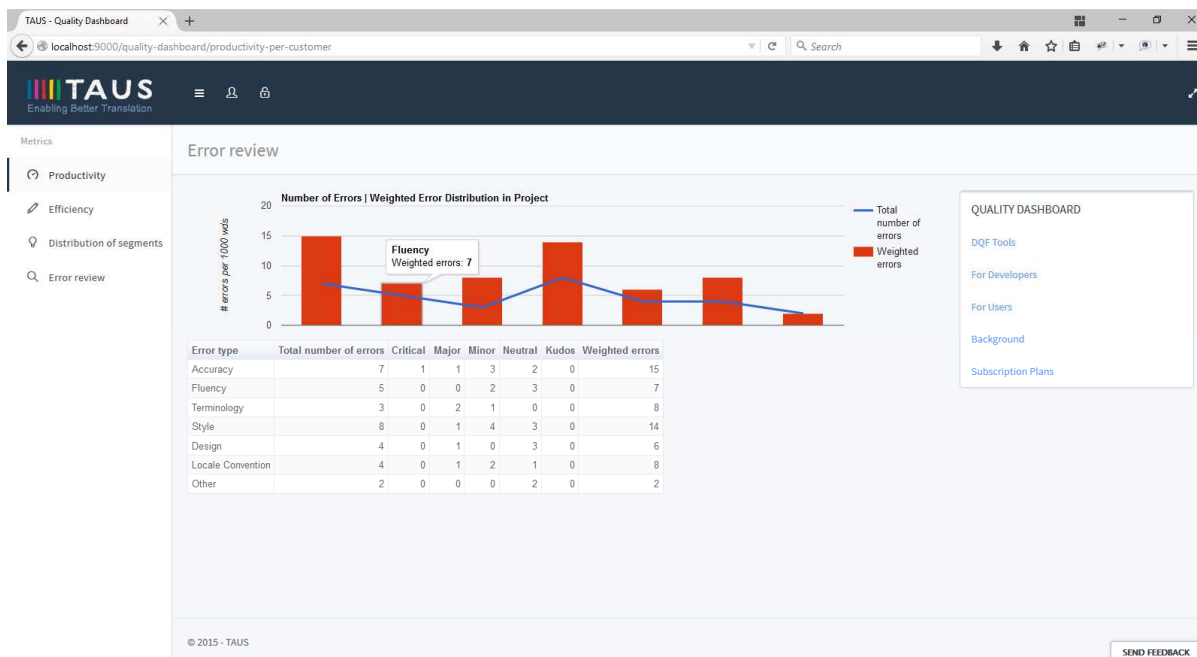


Figure 2 - Weighted error distribution

Project Managers may be interested to know how many and what kind of errors have been identified by each reviewer, as shown in Figure 3. This can be useful to compare different review styles and better understand the evaluation of e.g. in-country reviewer.

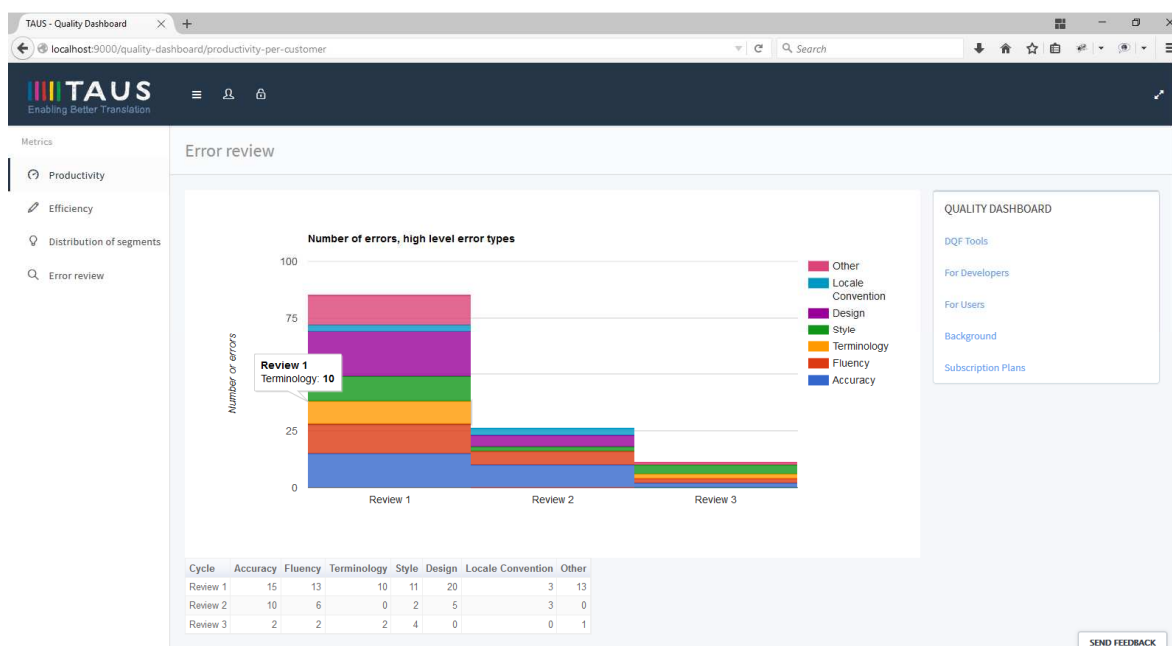


Figure 3 - Distribution of errors per task

5. Productivity and Efficiency

The Quality Dashboard provides productivity and efficiency metrics across content types, industries, processes used, technologies applied and by language pairs. Productivity is the throughput or speed expressed in the number of words per hour. Productivity tracking is

widely used for measuring the throughput of translators or quantifying the quality of MT engines by examining post-editing tasks. It helps evaluating which translation process is more appropriate and assessing the quality of the translation memory or the machine translation system in use.

Efficiency is a new score introduced by TAUS (Görög 2015a). The Efficiency Score is a composite indicator of translation productivity based on the words processed per hour and the edit distance. It calculates a weighted score, which gives a much more balanced and realistic insight in the performance of both human and technology resources than the commonly used productivity measurement (Görög 2015b). Using a similar procedure, additional attributes such as quality of the translated segments can be added to the Efficiency Score to reach higher precision.

While the productivity score is a good first performance indicator, the TAUS Efficiency Score gives both translators and managers a more reliable measurement, especially when used in combination with the filters for technology, process and content.

The Efficiency Score can be an absolute score calculated based on one given project or a relative score that is calculated using all the relevant data in the DQF database. It can be calculated using the two obligatory variables (core variables of words per hour and edit distance) or by adding some optional variables to the calculation to increase precision and credibility. It can be calculated to measure translator efficiency as well as CAT/TMS or MT engine efficiency.

6. Error Typology

A vast majority of providers and buyers of translation services manage their quality program with an error typology template. The LISA QA model and the SAE J2450 are among the two most commonly applied metrics for error category. TAUS has developed a more up-to-date version of these error typologies and made it available under DQF. The DQF error typology approach to quality evaluation involves the use of a list of error categories. The entire text or a sample thereof is evaluated by a qualified linguist who flags errors, applies penalties and establishes whether the content meets a pass threshold. This is a common type of evaluation in the translation sector. Although the error categories might vary, a benchmarking report by TAUS found that there was considerable similarity between the most commonly used typologies by over 20 companies (Language, Terminology, Accuracy and Style) and the types of errors. However, there is less agreement on the penalties to be applied or their severity levels.

In 2014, the German Research Center for Artificial Intelligence (DFKI) published the MQM (Multidimensional Quality Metrics) framework as part of the EU-funded QTLaunchPad project based on careful examination and extension of existing quality models (Lommel 2014). MQM is a framework for building task-specific translation metrics. It allows users to create custom metrics that can be used for various assessment purposes. By providing a master vocabulary of error types, users can describe metrics in a fully transparent fashion. MQM has been implemented in a variety of commercial and open-source tools.

Under the European funded project QT21, TAUS and DFKI have harmonized the DQF and MQM error typologies¹ into one DQF-MQM framework where the high-level branches match the six core DQF issue types (Figure 4). DQF's analytic method and the MQM hierarchy of translation quality issues have both been modified to share the same basic

¹ For more information about MQM, please visit <http://qt21.eu/mqm-definition>

structure. DQF will use a subset of the full MQM hierarchy based on the experience of TAUS members, while MQM will continue to maintain a broader set of issue types designed to capture and describe the full range of quality assessment metrics currently in use. Users of the DQF analytic method will be guaranteed to be compliant with MQM as well.

For each of the six main categories (Accuracy, Fluency, Design, Locale convention, Terminology, Style) there are a number of subcategories available for a more granular analysis of errors. For a complete list and description of the harmonized error categories (including the additional categories of ‘Verity’ and ‘Other’), please refer to the Appendix.

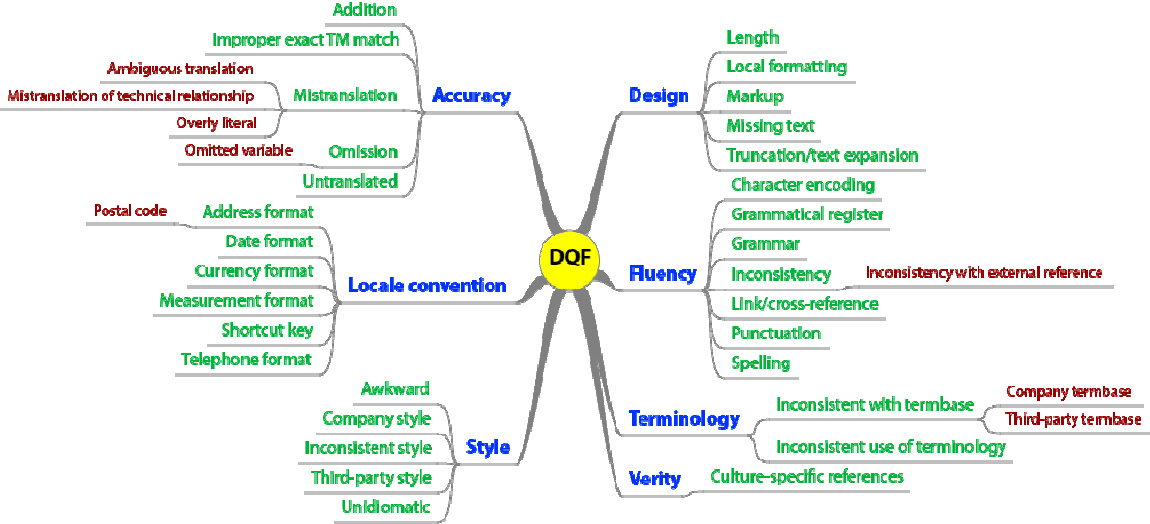


Figure 4 – The harmonized DQF-MQM error typology

The error typology approach is used to identify and classify errors in the text before delivery. Alternatively, error typology is employed to assess the performance of a vendor or identify mistakes in a machine translated document and guide their improvement. Based on the specific needs of the user, the error-typology can be more or less granular. A diagnostic evaluation to understand in detail the nature or cause of errors may require a more detailed error typology. The error typology should also be flexible enough to accommodate a customized selection of (sub-)categories.

Once the desired error typology has been selected, errors can typically be assigned to one of four severity levels: **critical**, **major**, **minor** and **neutral**. Each severity level is assigned a weight (penalty) which contributes to establishing the pass/fail outcome for the translation. The pass/fail threshold is flexible and depends on content type, end-user profile and perishability of the content. Different thresholds might be applied to different content types or target languages. Pass/fail thresholds can be set manually at project creation and penalties can be set on the Quality Dashboard based on error severity and error annotation to be performed at segment or sub-segment level. In the latter case, errors can be identified by highlighting the target text directly in the tool environment.

7. Conclusions and Further Work

The Quality Dashboard represents the natural evolution of the DQF tools. Thanks to the open API and the increasingly advanced reporting features we hope to encourage users to adopt the TAUS DQF and use the Quality Dashboard to measure translation performance and production and benchmark translation activities and quality turning traditional quality

evaluation into business intelligence.

The new quality review complements the productivity and efficiency measurements. For instance, the validity of the Efficiency Score can be improved if information on the quality of translated content is made part of the score. Furthermore, interesting conclusions can be drawn from productivity measurements of the review cycle(s) from the Quality Dashboard. Finally, translation productivity and quality can be correlated; post-editors, translators and reviewers can be profiled etc.

In later releases, additional features could be added such as content profiling to allow for automatic selection of error severities and pass/fail thresholds. Adequacy and fluency evaluation of each segment may also be integrated in the API to complement error annotation and offer an additional perspective on quality review. Sampling approaches in quality review also need further scoping to ensure reliable and comparable results.

References

Attila Görög. 2015a. *The TAUS Efficiency Score. Introducing a new score for measuring productivity*. Paper presented at the TAUS Quality Evaluation Summit, 28 May 2015, Dublin.

Attila Görög. 2015b. *Translation Productivity Revisited*. Blog article. <https://www.taus.net/blog/translation-productivity-revisited> (last visited 25 October 2015).

ArleLommel. 2014. *Multidimensional Quality Metrics (MQM) Definition*. <http://www.qt21.eu/mqm-definition/definition-2015-06-16.html> (last visited 25 October 2015).

Jaap van der Meer and Attila Görög. 2015. *Dynamic Quality Framework Report 2015*. https://evaluate.taus.net/index.php?option=com_rsfiles&layout=preview&tmpl=component&path=Reports%2FFree+Reports%2FDQFReport-TAUS2015.pdf (last opened 25 October 2015).

APPENDIX

Harmonized DQF-MQM Error Typology

ID	High-level error types	Granular error-types	Definition	Example
1	Accuracy		The target text does not accurately reflect the source text, allowing for any differences authorized by specifications.	Translating the Italian word 'canali' into English as 'canals' instead of 'channels'.
11		Addition	The target text includes text not present in the source.	A translation includes portions of another translation that were inadvertently pasted into the document.
12		Omission	Content is missing from the translation that is present in the source.	A paragraph present in the source is missing in the translation
13		Mistranslation	The target content does not accurately represent the source content.	A source text states that a medicine should not be administered in doses greater than 200 mg, but the translation states that it should be administered in doses greater than 200 mg (i.e., negation has been omitted).
14		Over-translation	The target text is more specific than the source text	The source text refers to a “boy” but is translated with a word that applies only to young boys rather than the more general term
15		Under-translation	The target text is less specific than the source text	The source text uses words that refer to a specific type of military officer but the target text refers to military officers in general
16		Untranslated	Content that should have been translated has been left untranslated.	A sentence in a Japanese document translated into English is left in Japanese.
17		Improper exact TM match	An translation is provided as an exact match from a translation memory (TM) system, but is actually incorrect.	A TM system returns “Press the Start button” as an exact (100%) match, when the proper translation should be “Press the Begin button”.
2	Fluency		Issues related to the form or content of a text, irrespective as to whether it is a translation or not.	A text has errors in it that prevent it from being understood.
21		Punctuation	Punctuation is used incorrectly (for the locale or style)	An English text uses a semicolon where a comma should be used.
22		Spelling	Issues related to spelling of words	The German word <i>Zustellung</i> is spelled <i>Zustetlugn</i> .
23		Grammar	Issues related to the grammar or syntax of the text, other than spelling and orthography.	An English text reads “The man was seeing <i>the his wife</i> .”

24		Grammatical register	The content uses the wrong grammatical register, such as using informal pronouns or verb forms when their formal counterparts are required.	A text used for a highly formal announcement uses the Norwegian <i>du</i> form instead of the expected <i>De</i> .
25		Inconsistency	The text shows internal inconsistency.	A text uses both “app.” and “approx.” for approximately.
26		Link/cross-reference	Links are inconsistent in the text	An HTML file contains numerous links to other HTML files; some have been updated to reflect the appropriate language version while some point to the source language version.
27		Character encoding		
3	Terminology			
31		Inconsistent with termbase	A term is used inconsistently with a specified termbase	A termbase specifies that the term <i>USB memory stick</i> should be used, but the text uses <i>USB flash drive</i> .
32		Inconsistent use of terminology	Terminology is used in an inconsistent manner within the text.	The text refers to a component as the “brake release lever”, “brake disengagement lever”, “manual brake release”, and “manual disengagement release”.
4	Style			
41		Awkward		
42		Company style	The text violates company/organization-specific style guidelines.	Company style states that passive sentences may not be used but the text uses passive sentences.
43		Inconsistent style	Style is inconsistent within a text	One part of a text is written in a light and “terse” style while other sections are written in a more wordy style.
44		Third-party style		
45		Unidiomatic		
5	Design			
			There is a problem relating to design aspects (vs. linguistic aspects) of the content.	A document is formatted incorrectly
51		Length	There is a significant discrepancy between the source and the target text lengths.	An English sentence is 253 characters long but its German translation is 51 characters long.
52		Local formatting	Issues related to local formatting (rather than to overall layout concerns)	A portion of the text displays a (non-systematic) formatting problem (e.g., a single heading is formatted incorrectly, even though other headings appear properly).
53		Markup	Issues related to “markup”	Markup is used incorrectly, resulting in

			(codes used to represent structure or formatting of text, also known as “tags”).	incorrect formatting.
54		Missing text	Existing text is missing in the final laid-out version	A translation is complete, but during DTP a text box was inadvertently moved off the page and so the translated text does not appear in a rendered PDF version.
55		Truncation/ text expansion	truncation-text-expansion	The German translation of an English string in a user interface runs off the edge of a dialogue box and cannot be read.
6	Locale convention		Characters are garbled due to incorrect application of an encoding.	A text document in UTF-8 encoding is opened as ISO Latin-1, resulting in all “upper ASCII” characters being garbled.
61		Address format	Content uses the wrong format for addresses.	An online form translated from English to Hindi requires a street number even though many addresses in India do not include a house number.
62		Date format	A text uses a date format inappropriate for its locale.	An English text has “2012-06-07” instead of the expected “06/07/2012.”
63		Currency format	Content uses the wrong format for currency.	A text dealing with business transactions from English into Hindi assumes that all currencies will be expressed in simple units, while the convention in India is to give such prices in lakh rupees (100,000 rupees)
64		Measurement format	A text uses a measurement format inappropriate for its locale.	A text in France uses feet and inches and Fahrenheit temperatures.
65		Shortcut key	A translated software product uses shortcuts that do not conform to locale expectations or that make no sense for the locale	A software product uses CTRL-S to save a file in Hungarian, rather than the appropriate CTRL-M (for <i>menteni</i>).
66		Telephone format	Content uses the wrong form for telephone numbers	A German text presents a telephone number in the format (xxx) xxx - xxxx instead of the expected 0xx followed by a group of digits separated into groups by spaces.
7	Verity		The text makes statements that contradict the world of the text	The text states that a feature is present on a certain model of automobile when in fact it is not available.
71		Culture-specific reference	Content inappropriately uses a culture-specific reference that will not be understandable to the intended audience	An English text refers to steps in a process as “First base”, “Second base”, and “Third base”, and to successful completion as a “Home run” and uses other metaphors from baseball. These prove difficult to translate and confuse the target audience in Germany.
8	Other		Any other issues	

memoQ 2015

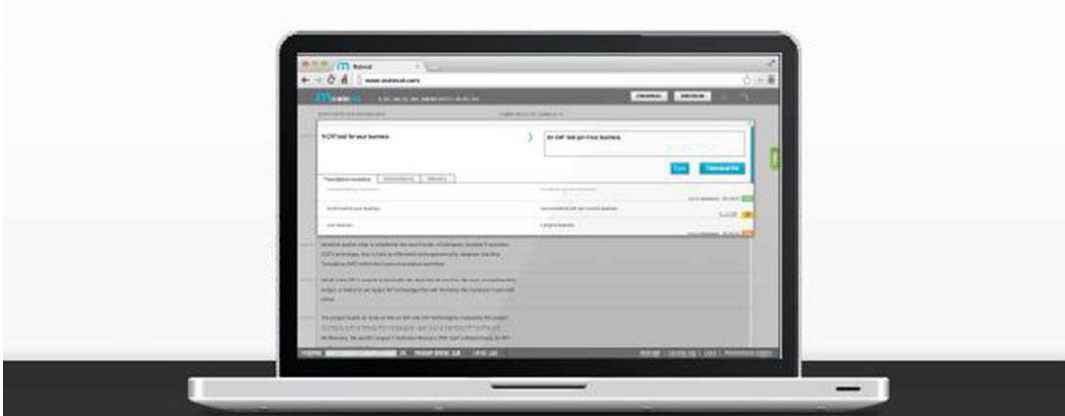
*“ MatchPatch is like driving
without your car with a GPS:
you can drive even without it,
but why would you make your
life more difficult? ”*

Ágnes Varga, PhD
memoQ Developer





More Matches for your Translations



Free and open source enterprise-level translation software

From 10% to 20% more matches than any other CAT tool

Increased privacy, no more files via email

A professional tool for language service providers and MT specialists



Collect data to set a fair rate for post-editor and improve MT quality



Real-time progress report and quality control for your translations



Online adaptation and quality estimation for MT systems based on Moses

Start translating www.matecat.com

Connect your Moses MT system via a set of open and easy to use API



The MateCat project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 287568.

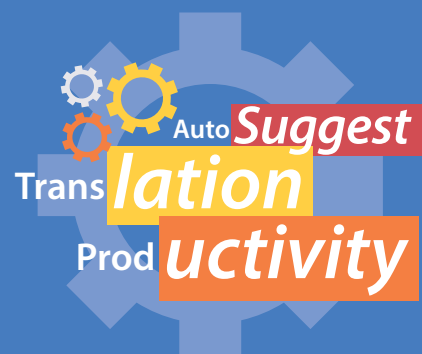
Translation - Terminology - Review & Quality - Query Management

Services - Software - Consulting

 eurocom®  quickTerm®  globalReview®  smartQuery®

Your Studio. Your Way.

Type less, translate more
with SDL AutoSuggest 2.0



Visit www.sdl.com/studio2015 or
www.translationzone.com/studio2015

SDL | Trados
Studio 2015



will organise

Translating and the Computer 38

17-18 November 2016

London (UK)

For information on next year's 38th Translating and the Computer conference, please check <http://translatingandthecomputer.com> and <http://asling.org> where information on calls for abstracts and posters, along with other information will be posted as it becomes available.