Deliverable 3.1
Harmonised Metric

Author(s): Arle Lommel (DFKI)
Attila Görög (TAUS)
Alan Melby (FIT standards committee and LTAC Global)
Hans Uszkoreit (DFKI)
Aljoscha Burchardt (DFKI)
Maja Popović (DFKI)

Dissemination Level: Public
Date: 2015-07-31
Copyright: © 2015 QT21 Consortium
Quality Translation 21
D3.1: Harmonised Metric

<table>
<thead>
<tr>
<th>Grant agreement no.</th>
<th>645452</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project acronym</td>
<td>QT21</td>
</tr>
<tr>
<td>Project full title</td>
<td>Quality Translation 21</td>
</tr>
<tr>
<td>Type of action</td>
<td>Research and Innovation Action</td>
</tr>
<tr>
<td>Coordinator</td>
<td>Prof. Josef van Genabith (DFKI)</td>
</tr>
<tr>
<td>Start date, duration</td>
<td>1 February 2015, 36 months</td>
</tr>
<tr>
<td>Dissemination level</td>
<td>Public</td>
</tr>
<tr>
<td>Contractual date of delivery</td>
<td>July 2015</td>
</tr>
<tr>
<td>Actual date of delivery</td>
<td>July 2015</td>
</tr>
<tr>
<td>Deliverable number</td>
<td>D3.1</td>
</tr>
<tr>
<td>Deliverable title</td>
<td>Harmonised Metric</td>
</tr>
<tr>
<td>Type</td>
<td>Report</td>
</tr>
<tr>
<td>Status and version</td>
<td>Version 6 final</td>
</tr>
<tr>
<td>Number of pages</td>
<td>33</td>
</tr>
<tr>
<td>Contributing partners</td>
<td>DFKI, TAUS</td>
</tr>
<tr>
<td>WP leader</td>
<td>FBK</td>
</tr>
<tr>
<td>Task leader</td>
<td>DFKI</td>
</tr>
<tr>
<td>Author(s)</td>
<td>Arle Lommel (DFKI) Attila Görög (TAUS) Alan Melby (FIT standards committee and LTAC Global) Hans Uszkoreit (DFKI) Aljoscha Burchardt (DFKI) Maja Popovic (DFKI)</td>
</tr>
<tr>
<td>EC project officer</td>
<td>Susan Fraser</td>
</tr>
</tbody>
</table>

The partners in QT21 are:

- Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI), Germany
- Rheinisch-Westfälische Technische Hochschule Aachen (RWTH), Germany
- Universiteit van Amsterdam (UvA), Netherlands
- Dublin City University (DCU), Ireland
- University of Edinburgh (UEDIN), United Kingdom
- Karlsruher Institut für Technologie (KIT), Germany
- Centre National de la Recherche Scientifique (CNRS), France
- Univerzita Karlova v Praze (CUNI), Czech Republic
- Fondazione Bruno Kessler (FBK), Italy
- University of Sheffield (USFD), United Kingdom
- TAUS b.v. (TAUS), Netherlands
- text & form GmbH (TAF), Germany
- TILDE SIA (TILDE), Latvia
- Hong Kong University of Science and Technology (HKUST), Hong Kong SAR of People's Republic of China

For copies of reports, updates on project activities, and other QT21-related information, contact:

DFKI GmbH
QT21
Prof. Stephan Busemann
Stuhlsatzenhausweg 3
Campus D3_2
D-66123 Saarbrücken, Germany
Phone: +49 (681) 85775 5286
Fax: +49 (681) 85775 5338

stephan.busemann@dfki.de

Copies of reports and other material can also be accessed via [http://www.qt21.eu](http://www.qt21.eu)
© 2015 QT21 Consortium

Page 2 of 33
1 Contents

1. Introduction and background .................................................. 4

2 Terminology .............................................................................. 7

3 Harmonising MQM and DQF ....................................................... 9
   3.1 Process .............................................................................. 9
   3.2 Challenges faced in the harmonisation ................................ 9
   3.3 Three new MQM dimensions ............................................. 11
      3.3.1 Terminology .............................................................. 11
      1.1.1 Locale convention .................................................... 12
      1.1.2 Style ....................................................................... 12
      1.2 Graphical representation ............................................. 12
      1.3 Scoring model ............................................................ 13
      1.4 Additional changes ..................................................... 14

2. The QT21 Metric for MT Evaluation ........................................... 15

3 Evaluation of reliability ............................................................... 18

A Annex A: Description of previous versions of MQM ................. 23

B Annex B: Description of DQF .................................................... 27
   B1. DQF in General ................................................................. 27
   B2. Quality Dashboard .......................................................... 27
   B3. Error-typology Evaluation .............................................. 27
      B.3.1 Background ............................................................... 27
      B.3.2 Scenarios .................................................................. 28
      B.3.3 Overview .................................................................. 28
      B.3.4 Evaluation Approach .............................................. 28
   B4. Error-typology in DQF tools ............................................. 29
   B5. Definitions for the main error categories ......................... 29
   B6. Definitions for the severity levels ................................... 29

C Annex C: Listing of issues types in DQF subset ....................... 32
1. Introduction and background

A major area of work in the QT21 project is (human) annotation of (machine translation) errors to assess quality and provide precise input into improvement efforts. Previous work on MT quality has relied primarily on the following methods:

1. Comparison of MT output to human reference translations (e.g., BLEU, NIST)
2. Analysis of post-editing effort or degree of change
3. Categorization of human post-editing operations (e.g., Hjerson, but also considerable anecdotal/observational work)

These approaches provide useful insights, but do not provide a detailed analytic picture of the precise kinds of errors made. In the case of reference-based approaches such as BLEU, only a numerical value with respect to a reference or set of reference translations is produced, with no indication of the precise differences. In the case of post-editing effort, a more direct understanding of quality is possible, but it also does not provide detailed insight into system's strengths and weaknesses. Categorization of post-editing operations, by contrast, does provide insight into specific changes that is very useful, but it is limited to describing what changed, but not why it changed or what was wrong.

In opposition to these methods, the evaluation of human translation has long used a "defect"-based approach in which specific errors are directly identified and categorized. Very often in such approaches a score is generated that identifies the deviation of the translation from a "defect-free" translation. This general approach is widely used by translation service providers using in-house checklists, but has also been standardized (or quasi-standardized) in the SAE J2450 specification (for assessing quality of automotive service translations) and the LISA QA Model (for software and document localization).

More recent work has recognized that translations serve many different purposes and that there is no single ideal translation for any given text, but rather a variety of different translations that serve different purposes. For example, a legal translation will have very different requirements in terms of accuracy and adherence to locale-specific norms than will the translation of an advertisement, a popular novel, or a television script.

In response to the limitations of previous methods and in order to help bring the evaluation of human and machine translation closer, the previous QTLaunchPad project developed the Multidimensional Quality Metrics (MQM), a flexible system that organizes many types of translation errors (called "issues" in MQM) into a hierarchy that supports multiple levels of granularity. The MQM hierarchy was based on an examination of a large variety of existing translation quality metrics that was intended to be comprehensive and detailed. It was also intended to be flexible: users could select a relevant subset of MQM and use it to perform task-specific evaluations according to the type and purpose of the translation.

Simultaneously with the development of MQM, the Translation Automation User Society (TAUS) developed the Dynamic Quality Framework (DQF). As one major


component, DQF contained an analytic error typology. The error typology consisted of six primary issue types with an elaborated list of subtypes (a two-layer hierarchy). In contrast to the MQM hierarchy, which seeks to be comprehensive and define the full range of issues/errors that may arise in translation quality, DQF was based on industry best practice, with a focus on the issues commonly checked by translation service providers\(^3\). As such, the DQF hierarchy was smaller and flatter.

Both MQM and DQF have been used and implemented in various tools. MQM is the default quality model in the open-source translate5 tool\(^4\) and is also used by XTM,\(^5\) a commercial tool and in a DFKI-built tool called the Scorecard.\(^6\) MQM was used extensively and tested in QTLaunchPad and has also been deployed in a project funded by the Caribbean Regional Information and Translation Institute (CRITI), based in Paramaribo, Suriname. It has also been used in pilot testing by Mozilla for evaluating the quality of international software builds. DQF has been released as a set of tools through the TAUS website, but is now available through a set of APIs and DQF is being built into tools from a variety of developers. As a result the two were increasingly seen as being in competition with one another.

MQM and DQF have been worked out at roughly the same time in different environments and communities to support the evaluation of translation quality of both human and machine translation (see also the respective sections below). It had been discussed for some time that it would be good to relate both to avoid confusion among the growing number of users. In the reviews of the QTLaunchPad project, however, the reviewers made it clear that any further funding in this area was predicated upon the harmonization of the MQM and DQF error typologies. The reviewers and the Commission maintained that the presence of two substantively similar error typologies with no clear differentiation in purpose would create confusion and prevent adoption of either one. As a result harmonization was written into QT21 as a requirement for the project since the main developers (DFKI for MQM and TAUS for DQF) are partners in the project.

In this context, “harmonization” is to be understood as the comparison of the two frameworks, followed by the development of a shared error typology used by both frameworks. It does not mean that MQM and DQF unify or that they must be identical in every respect. For example, if the DQF error typology were a subset of the MQM typology, this would be acceptable since compatibility would be achieved.

Overall, the harmonisation of both resources went surprisingly well and posed few difficulties. Both partners made constructive changes and adaptations and the resulting harmonized hierarchy and shared derived metric, described in this Deliverable, is a much stronger specification than either was alone. This result shows that the very challenging task of evaluation of translation seems to have reached a level of maturity so that the different solutions can interoperate fairly well. As a result of the harmonisation, this Deliverable presents a description of the harmonized metric.

It also presents a preliminary evaluation of inter-annotator agreement (IAA) based on annotation by a set of fully trained expert annotators using a version of the MT

---

\(^3\) DQF also provides a number of other assessment methods that are independent of error typologies and are not addressed in this deliverable.

\(^4\) [http://translate5.net](http://translate5.net)

\(^5\) [http://www.xml-intl.com](http://www.xml-intl.com)

\(^6\) Source code available from [https://github.com/multidimensionalquality/qt21-scorecard](https://github.com/multidimensionalquality/qt21-scorecard)
assessment metric developed in the QTLaunchPad project that has now been updated to reflect the results of this harmonization. These preliminary results indicate that the new metric and training materials do deliver higher agreement than was found in previous annotation rounds in QTLaunchPad. While additional work remains to be done to fully determine IAA, these initial results are a positive indicator.
2 Terminology

The following terms apply to this document:

Dimension
The highest level of issue type in MQM. Dimensions have no parent issues and represent broad areas in which other issues are categorized. Dimensions in the revised MQM hierarchy presented in this document are: Accuracy, Design, Fluency, Internationalization, Locale convention, Style, Terminology, and Verity. In addition there is a Compatibility dimension that contains deprecated issues from legacy metrics (use is strongly discouraged) and Other, which is used for items that cannot be otherwise classified.

Error
A confirmed problem in a text. All errors are considered issues in MQM-compliant metrics, but not all issues are errors (since issues may be unconfirmed).

Error typology
Synonym of issue type hierarchy, used in DQF.

Issue
A potential problem detected in a text. Issues are categorized by their issue type. An issue may or may not be an error. For example, if an automatic process determines that a term is present in a source segment but not in the corresponding target, it has found an issue. A human reviewer might then find that the translation is accurate because the target segment used an unambiguous pronoun to refer to the concept, in which case the issue is not an error.

Issue type
A descriptive kind of an issue. For example, if a text exhibits a word order error, it can be classified using the Word order issue type. In this document the names of issue types are written in the Courier font, e.g., Word order, to distinguish them from running text.

Issue type hierarchy
Issue types are arranged in a hierarchy, with children of issues representing an is a type of relationship. Note that child nodes are not comprehensive of their parent node (i.e., they do not define every possible type of their parent), but instead represent important/common subtypes. For example, the issue type Word order is a type of Grammar, which is a type of Fluency. This relationship allows MQM and the DQF subset of MQM to deal with multiple levels of granularity, from very fine to very coarse.

Severity level
An indicator of the importance of an issue with an accompanying numerical representation. For example, MQM distinguishes between major and minor errors, where major errors impact understanding and minor errors do not.
Weight
A numeric representation of the importance of an issue type in a specific metric. For example, a metric might assign a weight of 2.0 to Terminology and 1.0 to Style, thus stating that Terminology is counted twice as important as Style for the relevant translation specifications.
3 Harmonising MQM and DQF

Note: Much of this section addresses details of the harmonisation process that require knowledge of MQM and DQF. Readers who are not familiar with these specifications are encouraged to consult Annexes A and B.

3.1 Process

Harmonization was conducted through weekly online meetings and email exchange starting in February 2015 and extending through early June 2015. This discussion primarily involved Attila Görög (TAUS), Arle Lommel (DFKI), and Alan Melby (member of the FIT Council and President of LTAC Global). The harmonization involved a systematic examination of the error hierarchies of both DQF and MQM and discussion about terminology, principles, and other factors. Most of the work focused on the issue-type hierarchies as they presented the most differences. (Principles and terminology were agreed upon quite early on.)

3.2 Challenges faced in the harmonisation

Although the MQM and the DQF error typologies are broadly similar, harmonisation on the level of details was a non-trivial task because hierarchy is vital to MQM, but comparatively less important to DQF. The following items and observations arose during harmonisation:

- The “Localization specific” and “Translation specific” branches in DQF were groupings for convenience and were never selected as issue types. Accordingly these could be omitted from the merger hierarchy.
- Although the DQF issue-type selection had only six “active” error types—Adequacy, Linguistic, Terminology, Style, Country Standards, and Layout—(plus four additional items that did not fit into MQM, as discussed in the next point), the examples in the DQF documentation in fact defined a deeper hierarchy. When this deeper hierarchy was considered, MQM could not represent the DQF structure without modification. As a result the structures of both MQM and DQF would need to be adjusted.
- DQF used the term Adequacy, which is widely used in MT evaluation, where MQM used Accuracy. Although often treated as equivalent, the relationship between these two is not entirely simple or clear, and requires further explication and examination.
- The DQF items query implementation, client edit, repeat, and kudos do not describe error types, but rather address other issues important in quality, such as marking changes made in response to queries to the content creator, edits made by the client, repeated errors, and credit given for particularly good translation solutions. These items cannot be represented directly in MQM.
- The definitions of some issues did not align (e.g., DQF style included “Ambiguous translation”, which was not included in MQM Style (it is, instead

---

a sister category to it). Even though the daughter issues in DQF were descriptive, the scope of the issues present in each system had to match if the systems were to be truly harmonised.

In order to resolve these issues, a series of regular meetings was held over a period of five months to discuss issues and reach compromises. During this time, the following general approach was taken:

- In cases where the same issue appeared in DQF and MQM but with different names, DQF adopted the MQM name.
- In cases where the structures of the two specifications differed, MQM would either be altered to reflect the structure of DQF or the two would both be modified at the same time to arrive at a compromise position.
- Any examples within DQF issue types that were not already in MQM would be added to MQM.

As MQM required more substantive changes to accommodate DQF than DQF did to accommodate MQM’s structure, the following description is written primarily from the perspective of MQM.

The primary changes made to MQM to accommodate DQF are the following:

- Three new dimensions were added to MQM: Terminology, Style, and Locale convention. These dimensions took issues from other dimensions, as noted below, and doing this allowed the DQF structure to be represented much more closely in MQM. (See Section 3.3 below for a discussion of this change.)
- Within the MQM Accuracy dimension:
  - Terminology was removed from Mistranslation and made into a dimension of its own.
  - Mistranslation gained a new daughter Mistranslation of technical relationship (from DQF). This issue type is used for cases of technical content in which text is translated in a way that appears correct and plausible (i.e., the target text could be a valid translation of the source text), but which nevertheless fails to accurately render a technical concept. This issue is commonly encountered when a translator is not familiar with a technical concept and the source text is ambiguous when the reader does not know the technical domain (but generally will be unambiguous for a reader who does know it).
- Within the MQM Fluency dimension:
  - The distinction between Mechanical and Content was completely eliminated and all children were moved directly under Fluency. While theoretically nice, in practical terms it caused problems.
  - Under issue type Inconsistency:
    - Moved Terminological inconsistency to the MQM Terminology top level dimension
    - Added Inconsistency with external reference for compatibility with DQF
  - Eliminated Monolingual terminology: the category was not widely used and was confusing to people. All terminology-related issues are now in the Terminology dimension and have been restructured for greater clarity.
moved all Locale convention issue types to the new dimension Locale convention, which corresponds to DQF’s country standards.

- A new issue type, Grammatical register, was added to Fluency and replaces Register in the MQM Core. It deals specifically with grammatical markers of formality (e.g., German du vs. Sie).
- Register was moved to be a child in the new Style dimension, and takes on a slightly narrower meaning: it is now the equivalent of the DQF Tone error type, and deals with stylistic correlates to register (e.g., formal/informal style).

- Within Locale convention:
  - Added Culture-specific reference (from DQF)
  - A number of additional types were added to match the TAUS list.

- Style guide and Stylistics were recombined (they had been split to deal with the Mechanical/Content division) back into Style, which was made a top level dimension.

- The following additional changes were made in Style:
  - Added Awkward
  - Replaced Style guide with Company style and Third-party style (a style guide would always be one of these types)
  - Added Inconsistent style as a new issue type within Style.

- Moved Unidiomatic from Content to Style

In addition to the changes noted above, there were a number of other minor changes made to improve compatibility with a proposed standard for Language Quality Assurance (LQA) currently under development in ASTM F43\(^8\) and to address other issues that had been raised in the previous months since the end of QTLaunchPad.

### 3.3 Three new MQM dimensions

The decision to add new dimensions makes it so that the default six error types in DQF correspond exactly to six of the eight MQM dimensions. As a result DQF serves as a very high-level MQM metric that covers all of the dimensions except Verity\(^9\) and Internationalization, which does not apply to translated texts but instead only to source texts.

#### 3.3.1 Terminology

The decision to create a separate Terminology dimension addressed a long-standing confusion in MQM: terminology-related issues were split into various dimensions and the primary Terminology issue was nested as a subtype of Mistranslation. Treating adherence to a termbase or to domain conventions as a Mistranslation issue had resulted in consistent confusion in testing during QTLaunchPad. It also ran counter to general industry practices that often assess compliance to terminology guidelines in a separate process. In addition,


\(^9\) Verity is included in the full DQF subset to cover Culture-specific references, but was not one of the basic six issue types in DQF and so is not included unless the full DQF hierarchy is used.
Terminology is the only issue type assessed in every metric that was evaluated in creating MQM, so nesting it deep in the hierarchy was problematic.

By contrast, DQF had always treated Terminology as a high-level issue, and nesting it downward would have been problematic for DQF. Accordingly the decision was made to move it up to be its own dimension and to move all terminology-related issues from other branches.

This change enabled MQM to remove another confusing distinction: it had originally distinguished between bilingual terminology (i.e., the relationship between terms in source and target language) and monolingual terminology (i.e., the appropriateness of terms without regard to whether they were translated or not). MQM had also treated “inconsistent use of terminology” under Inconsistency. Moving these issue all to one place eliminated the need to consider which dimension a terminology problem might be assigned to and made it easier to find the appropriate issue. It also made the intention clearer than was possible when Terminology was treated as a subtype of Mistranslation.

1.1.1 Locale convention

The issues under Locale convention relate to the locale-specific formatting of content. Their position had always been unclear and they had briefly been placed under Verity. However, seeing that they were a high-level category in DQF led the team to realize that they should also be a high-level category in MQM (and thus in the harmonised hierarchy). The dimension was expanded to include types from MQM, with additional types added to correspond to an expansion of the Internationalization dimension that happened to take place at the same time.

1.1.2 Style

The issues under Style were moved from various places to this new dimension. It covers both formally defined style (such as is found in a formal style guide like Chicago Manual of Style) and style in the broader sense of linguistic artistry and feeling. Register, which had formerly been a sibling of Style within Fluency, is now included under Style, where it corresponds to the DQF Tone category. To deal with the issue of grammatical register (e.g., Sie vs. du in German), we added a new issue type, Grammatical register, under Fluency. Splitting register into two categories allows the stylistic aspects of register (e.g., such as a formal style of writing) to be separated from the narrow issue of grammatical register/formality (which are more closely related to general fluency).

1.2 Graphical representation

The results of the harmonization are described online at http://qt21.eu/mqm-definition. Due to space limitations, the full MQM definition is not included here. However, the MQM-compatible version\(^5\) of DQF is presented in Figure 1:

---

\(^5\) DQF is a compliant subset of MQM, but it is possible to create MQM subsets that are not compatible with DQF by using MQM issue types that do not correspond to DQF. DQF is a simplified subset of MQM that excludes less-common issues. In most cases an MQM subset could be converted to DQF by using appropriate issue types in the MQM hierarchy that correspond to DQF issue types. However, if the Internationalization dimension is used, an MQM metric will not be DQF compatible since there is no corresponding portion of DQF (note, however, that Internationalization is not used for assessing translated texts in any event, since it focuses on errors in source content).
1.3 Scoring model

An additional major change was made to the MQM default scoring model to accommodate features of DQF that are not issue types per se, but rather ways of marking changes that need to be made that are not considered errors in the translation. These DQF items are query implementation, client edit, and repeat. Since MQM marks issue types, there was no easy way to handle these items in MQM. The solution that was arrived at was to add a fourth default error severity level (besides minor, major, and critical) called none. Issues marked with this severity level do not count against any quality score for the translation, but processes may require them to be addressed. To address the DQF items, issues are to be marked with the severity level none and categorized as normal using MQM issue
types. Then notes can be appended explaining the nature of the issue per the DQF typology.

This shift does not address the DQF kudos category, which is used to add points to a translation. What to do with this item (which is not widely implemented) remains an open point of discussion, but does not affect any QT21 tasks since it would not be used in the QT21 error annotation.

1.4 Additional changes

At the same time as the harmonization was taking place, QT21 became aware of additional issues and of another quality assessment effort (LQA\textsuperscript{11}) going on in ASTM F43\textsuperscript{12}. As a result additional changes were made. These changes are described in the MQM specification in the list of changes. The most notable changes were the addition of two new subtypes of Accuracy, Over-translation and Under-translation. These, respectively, refer to cases in which a translation is accurate but narrower in meaning than the source and cases in which a translation is accurate but broader in meaning. Such differences are inevitable in some cases (e.g., one language makes meaning distinctions that another does not) but also frequently are a marker of cases in which a translator has not fully understood a source text and so is too narrow or insufficiently precise.


\textsuperscript{12} http://www.astm.org/COMMITTEE/F43.htm
2. The QT21 Metric for MT Evaluation

Although DQF and MQM have been harmonized, the quality evaluation tasks within QT21 require a degree of granularity in analysis that goes beyond the DQF subset, which was designed for general industrial use rather than the detailed analysis needed in a research context (and adding these issues to DQF would have a negative impact in an industrial context by raising complexity). At the same time, any subset of MQM used must be compatible with DQF (i.e., it must not contradict it and it must be reducible to the DQF subset) if QT21 is to truly use a harmonized metric.

The decision was made to update the MQM subset used for MT evaluation in the QTLaunchPad project. Taking this approach has two advantages:

1. The project team and some potential annotators are already familiar with the existing set, a factor that would reduce training effort.
2. Any replacement issue set would need to be substantially similar to the existing metric to meet project requirements, so developing an entirely new metric would serve little purpose.

At the same time, the metric previously used in QTLaunchPad would need to be updated in any event to match the newest MQM version.

The previous version of this metric’s issue types is as follows (Figure 2):

![Figure 2 Original QTLaunchPad issue types](image)

This metric was subsequently modified for use in the QTLeap project by adding in Terminology as a child of Mistranslation. The proposed metric would retain Terminology as a top-level dimension and also add Locale convention as a dimension and Grammatical register as an issue under Fluency. This revised set of issues can be visualized as shown in Figure 3. In this figure, items in italics with a light blue background correspond to DQF issues. Additional issues not present in DQF are treated as types of their parent for purposes of DQF compatibility.

(Note that this metric includes three issue types—Extraneous, Incorrect, and Missing—as custom subtypes of Function words. These are not generally used in MQM, but can be considered explanatory subtypes of Function words. They are not included in general MQM because they are too specific to QT21 requirements.)
The mapping of issues is as follows:

<table>
<thead>
<tr>
<th>Diagnostic Metric Issue Type</th>
<th>DQF High-Level Type</th>
<th>DQF Lower-Level Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>Accuracy</td>
<td>Accuracy</td>
</tr>
<tr>
<td>Addition</td>
<td>Accuracy</td>
<td>Accuracy</td>
</tr>
<tr>
<td>Mistranslation</td>
<td>Accuracy</td>
<td>Accuracy</td>
</tr>
<tr>
<td>Omission</td>
<td>Accuracy</td>
<td>Accuracy</td>
</tr>
<tr>
<td>Untranslated</td>
<td>Accuracy</td>
<td>Accuracy</td>
</tr>
<tr>
<td>Fluency</td>
<td>Fluency</td>
<td>Fluency</td>
</tr>
<tr>
<td>Grammar</td>
<td>Fluency</td>
<td>Fluency</td>
</tr>
<tr>
<td>Function words</td>
<td>Fluency</td>
<td>Fluency</td>
</tr>
<tr>
<td>Extraneous</td>
<td>Fluency</td>
<td>Fluency</td>
</tr>
<tr>
<td>Incorrect</td>
<td>Fluency</td>
<td>Fluency</td>
</tr>
<tr>
<td>Missing</td>
<td>Fluency</td>
<td>Fluency</td>
</tr>
<tr>
<td>Word form</td>
<td>Fluency</td>
<td>Fluency</td>
</tr>
<tr>
<td>Agreement</td>
<td>Fluency</td>
<td>Fluency</td>
</tr>
<tr>
<td>Part of speech</td>
<td>Fluency</td>
<td>Fluency</td>
</tr>
<tr>
<td>Tense/aspect/mood</td>
<td>Fluency</td>
<td>Fluency</td>
</tr>
<tr>
<td>Word order</td>
<td>Fluency</td>
<td>Fluency</td>
</tr>
<tr>
<td>Grammatical register</td>
<td>Fluency</td>
<td>Fluency</td>
</tr>
<tr>
<td>Spelling</td>
<td>Fluency</td>
<td>Fluency</td>
</tr>
<tr>
<td>Typography</td>
<td>Fluency</td>
<td>Fluency</td>
</tr>
<tr>
<td>Unintelligible(^{13})</td>
<td>Fluency</td>
<td>Fluency</td>
</tr>
<tr>
<td>Locale convention</td>
<td>Locale convention</td>
<td>Locale convention</td>
</tr>
<tr>
<td>Terminology</td>
<td>Terminology</td>
<td>Terminology</td>
</tr>
</tbody>
</table>

Table 1. Mapping from MQM issue types to DQF

As can be seen in Table 1, the MQM diagnostic metric is much finer grained than DQF (either with the six basic categories or with the subtypes). However, using the

\(^{13}\) Note that some of the issues in the metric go beyond the categories in DQF, as is the case with Unintelligible. In such cases these issues can be treated as belonging to the next highest issue type in the MQM hierarchy: in this case the MQM issue type Unintelligible would be treated as general Fluency in DQF, since DQF does not have a specific category for this type. Because this metric can be reduced to a clean subset of MQM, compatibility is maintained. (If the metric used issues that could not be reduced to DQF, it would be incompatible, but that is not the case in this metric.)
higher-level categorizations from DQF will help in establishing at what level annotators can agree on issue types since they may disagree on fine-grained detail. For example, they might disagree on whether a given problem has to do with Word form or Grammatical register and still agree that it is a Grammar problem. Since the MT diagnostic metric can be easily reduced to either the full depth of DQF typology or the smaller set of four issues shown in the second column above, DQF provides a useful abstraction to see where evaluators are in fundamental agreement about the nature of a problem, even if they disagree on the finer-grained details.
3. Evaluation of reliability

The reliability (i.e., the agreement between annotators performing the same task) of MQM for annotation was a point of concern in the QTLaunchPad project since it did not reach bounds considered acceptable to the reviewers. The measure of reliability used in that project was inter-annotator agreement (IAA), an indication of the likelihood that reviewers would agree in their annotations.

When an initial version of the metric described in this deliverable was used in late 2013 and early 2014 in the QTLaunchPad project, it was found to have relatively low IAA (an F-score averaging 33.0 across four language pairs with 36.5 for EN>DE). Subsequent training and development increased the F-scores to 48.8 (average) and 47.8 (EN>DE), a significant increase.\textsuperscript{14} However, these scores fell below desired levels (for many purposes a score of 90 would be considered a minimal level of reliability).

Although these values thus seemed quite low, they actually compared quite favourably with other tasks, such as the WMT ranking tasks, which typically did not exceed an equivalent value of approximately 30. Nevertheless, improvement of reliability was considered a highly desirable outcome.

In order to determine whether trained evaluators could improve upon the previous results, we considered sentence-level F-scores once again\textsuperscript{15} and ran a test with the annotation of 300 segments of WMT data going from English to German using translate5.\textsuperscript{16} The data set from which these segments were drawn had previously been selected in the QTLaunchPad project as examples of “near miss” translations, i.e., translations that could be used with relatively few errors. This selection criterion was used to avoid the scenario in which annotators would be asked to annotate sentences that had so many errors that individual errors could not be discerned or separated from other errors. In addition, using these segments would permit direct comparison with IAA with QTLaunchPad. All of the three annotators involved had previous experience with annotation in QTLaunchPad, but had not actively done annotation work in several months.

The annotators each found an average of 633 issues in the 300 segments, or an average of 2.11 errors/segment, consistent with the previous classification of the segments as “almost good.” This number was sufficient to calculate sentence-level agreement for the issue types evaluated, but not sufficient to evaluate all individual issue types, as will be explained below.

One difference between this test and previous tests was that the annotators were asked to simultaneously post-edit the segments they were evaluating so that the

\textsuperscript{14} QTLaunchPad Deliverable D1.3.1, Barriers for High-Quality Machine Translation, pp 4–6 (http://www.qt21.eu/launchpad/system/files/deliverables/QTLP-Deliverable-1_3_1-v2.5.pdf)

\textsuperscript{15} F-scores were calculated using a python script created by Maja Popović. This script can be downloaded from http://qt21.eu/downloads/F-score-script.zip. To use this script, a list of issue types found in each segment was extracted (no span-level information was retained) and separate files were saved for each annotator. One annotator was treated as the reference and the other as the hypothesis (F-scores are calculated in pair-wise combinations). The results are obtained with the following command (assuming the input files are called A.txt and B.txt):

\texttt{python rgbF.py -R A.txt -H B.txt -n 1}

\textsuperscript{16} Annotation guidelines used in this task are available in QT21 Deliverable D5.5, Data management plan. All annotators were familiar with the translate5 environment and had previously been trained on the tool.
results could be interpreted against their posted-edited version that reflects the way they would solve the problems they found in the segments. The results of the post-edits have not yet been correlated with annotation due to personnel changes at DFKI: these results will be released in future deliverables when more data are available.

The results were highly encouraging. The average F-score in this test was 52.45 when using the full hierarchy of the metric described in the previous section. This result shows that, at the sentence level, the newer version of the metric shows improved IAA.

In addition we calculated the F-score using just the top-level DQF categories (i.e., all issues were reduced to the top-level dimensions of Accuracy, Fluency, Locale convention, and Terminology)\(^\text{17}\) to see if, even when annotators disagreed about the precise nature of an issue, they agreed on the broad conceptual type of issue. Here the average F-score was 68.4. This result contrasts very favourably with work done in QTLaunchPad where moving up the hierarchy did not have a significant impact, perhaps indicating that the current split into more dimensions is a better cognitive match for how annotators understand the issues they encounter.

Table 2 summarizes these changes. As can be seen, the creation of the QTLP Round 2 metric (which is very similar to the metric used in QT21) resulted in a substantial increase in IAA, while the current metric results in a continued improvement over QTLP results. Using the simplified four-issue metric derived from DQF results in

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>EN-&gt;DE</td>
<td>36.5</td>
<td>47.8 (+11.3)</td>
<td>52.5 (+26.0 over QTLP1)</td>
<td>68.4 (+16.9 over full)</td>
</tr>
<tr>
<td>Average (EN-&gt;DE, EN-&gt;ES)</td>
<td>33.0</td>
<td>48.8</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

Table 2. Change in IAA between QTLaunchPad and QT21.

Note that 100% agreement in all details is highly unlikely when using MQM, for the following reasons:

- Different annotators may disagree as to whether certain phenomena are acceptable or not, depending on dialect and individual linguistic preferences. For example, in English, one individual might consider a so-called “split infinitive” to be a Grammar error while another finds it perfectly acceptable. In QTLaunchPad it was found that such cases were not uncommon: some reviewers were simply more tolerant than others.
- Many phenomena can be analysed in multiple, equally valid fashions. If the annotators have different solutions to problematic portions in mind, they may categorize the problems in different fashions. For example, one might identify

\(^{17}\) For example, in this score the following issues would all be counted as Accuracy: Accuracy, Addition, Mistranslation, Omission, and Untranslated. A similar procedure was applied for Fluency. As no subtypes for Locale convention or Terminology were used, these dimensions were not changed at all.
a portion as a Mistranslation, while another might see a way to move words around (word order) that results in a valid translation: they would thus see the same passage and problem, but disagree on its nature.

Based on experience in QTLaunchPad, the project team had estimated that IAA might not exceed 70% due to these differences (i.e., legitimate disagreement might approach 30% of instances). Using the DQF high-level categories, then, it seems that IAA may be approaching its maximum value.

Further examination shows that agreement varies considerably by issue type. However, some issues occurred so infrequently that drawing significant conclusions about agreement based on them is impossible. The percentage of each issue (out of all issues) is given in Table 3 (below):

<table>
<thead>
<tr>
<th>Issue type</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mistranslation</td>
<td>40.9%</td>
</tr>
<tr>
<td>Agreement</td>
<td>10.1%</td>
</tr>
<tr>
<td>Incorrect function word</td>
<td>9.6%</td>
</tr>
<tr>
<td>Omission</td>
<td>7.1%</td>
</tr>
<tr>
<td>Tense/aspect/mood</td>
<td>6.6%</td>
</tr>
<tr>
<td>Missing function word</td>
<td>6.4%</td>
</tr>
<tr>
<td>Typography</td>
<td>6.0%</td>
</tr>
<tr>
<td>Extraneous function word</td>
<td>3.6%</td>
</tr>
<tr>
<td>Spelling</td>
<td>2.3%</td>
</tr>
<tr>
<td>Addition</td>
<td>1.9%</td>
</tr>
<tr>
<td>Part of speech</td>
<td>1.1%</td>
</tr>
<tr>
<td>Untranslated</td>
<td>1.0%</td>
</tr>
<tr>
<td>Word form</td>
<td>1.0%</td>
</tr>
<tr>
<td>Unintelligible</td>
<td>0.9%</td>
</tr>
<tr>
<td>Grammatical register</td>
<td>0.4%</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.4%</td>
</tr>
<tr>
<td>Locale convention</td>
<td>0.4%</td>
</tr>
<tr>
<td>Terminology</td>
<td>0.3%</td>
</tr>
<tr>
<td>Grammar</td>
<td>0.1%</td>
</tr>
<tr>
<td>Function words</td>
<td>0.0%</td>
</tr>
<tr>
<td>Fluency</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

Table 3. Distribution of issues assigned in inter-annotator agreement evaluation round.

As can be seen, Mistranslation and Agreement together comprise the majority of all issues found and the top 8 issue types (bolded) account for over 90% of all issues found. Extraneous function word, averaged 20 instances per reviewer, while others were very infrequent, making it difficult to assess their reliability over 300 segments. This distribution is consistent with previous annotation exercises and the small numbers of a number of issues comprise a classical "long tail" distribution. For issues in this long tail, IAA figures per issue type are therefore inconclusive.

With these provisos, Table 4 presents a simple measure of agreement for each issue type as well as for branches in the hierarchy (e.g., counting all types of Function word issues as their parent issue or all children of Grammar as Grammar). Note that F-scores cannot be used in this per-issue context because they will be artificially high in instances where only a few occurrences of an issue type are present in the corpus.
since agreement that an issue does not apply is counted as agreement, thus drowning the agreement in positives in a signal from negatives.

<table>
<thead>
<tr>
<th>Issue type</th>
<th>% agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>50</td>
</tr>
<tr>
<td>Addition</td>
<td>81</td>
</tr>
<tr>
<td>Mistranslation</td>
<td>66</td>
</tr>
<tr>
<td>Omission</td>
<td>38</td>
</tr>
<tr>
<td>Untranslated</td>
<td>0</td>
</tr>
<tr>
<td><strong>Accuracy subtotal</strong></td>
<td><strong>61</strong></td>
</tr>
<tr>
<td>Fluency</td>
<td>0</td>
</tr>
<tr>
<td>Grammar</td>
<td>0</td>
</tr>
<tr>
<td>Function words</td>
<td>0</td>
</tr>
<tr>
<td>Extraneous</td>
<td>20</td>
</tr>
<tr>
<td>Incorrect</td>
<td>44</td>
</tr>
<tr>
<td>Missing</td>
<td>47</td>
</tr>
<tr>
<td><strong>Function words subtotal</strong></td>
<td><strong>40</strong></td>
</tr>
<tr>
<td>Word form</td>
<td>0</td>
</tr>
<tr>
<td>Agreement</td>
<td>80</td>
</tr>
<tr>
<td>Part of speech</td>
<td>0</td>
</tr>
<tr>
<td>Tense/aspect/mood</td>
<td>0</td>
</tr>
<tr>
<td><strong>Word form subtotal</strong></td>
<td><strong>62</strong></td>
</tr>
<tr>
<td>Word order</td>
<td>61</td>
</tr>
<tr>
<td><strong>Grammar subtotal</strong></td>
<td><strong>53</strong></td>
</tr>
<tr>
<td>Grammatical register</td>
<td>0</td>
</tr>
<tr>
<td>Spelling</td>
<td>23</td>
</tr>
<tr>
<td>Typography</td>
<td>44</td>
</tr>
<tr>
<td>Unintelligible</td>
<td>0</td>
</tr>
<tr>
<td><strong>Fluency subtotal</strong></td>
<td><strong>59</strong></td>
</tr>
<tr>
<td>Locale convention</td>
<td>50</td>
</tr>
<tr>
<td>Terminology</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 4. Likelihood of annotators agreeing on individual issues.

To understand how these figures are calculated, consider a case in which two annotators found the following issues for a segment:

<table>
<thead>
<tr>
<th>Annotator A</th>
<th>Annotator B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mistranslation, Mistranslation, Spelling</td>
<td>Mistranslation, Spelling</td>
</tr>
</tbody>
</table>

In this case, Annotator A and B both show 100% agreement on Spelling. For Mistranslation, Annotator A shows 50% agreement with Annotator B while B shows 100% agreement with A (since A found everything B did). Overall, using this model, B shows 100% agreement with A and A shows 67% agreement with B. And
the average agreement for the segment is 83.3%. Because comparisons are not reversible (i.e., comparing A to B will give different results than comparing B to A), the resulting figures must be averaged.

Unlike with F-scores, this method does not directly penalize cases where one annotator has more instances than the one being evaluated (although it will penalize the score when the calculation is performed for the other annotator and then averaged), nor does it address the probability of random agreement. However, it allows easy comparison of the likelihood that issues found in one translation will be found in the others against which it is compared.

A score of 0 for an issue can either mean that there was no agreement between annotators or that the issue was not annotated at all by any annotators (in the case of Fluency and Function words). All figures are averages across all pairwise comparisons.

The numbers show that different issues show very different profiles in terms of IAA. These results, combined with figures for the relative frequency of individual issues will be used to improve training materials for future annotation rounds by helping to identify issues that show particular difficulty for annotators.

One striking result is the low IAA for Terminology: this result emerged because only one annotator used this category. As WMT data is from the news domain (which is not a proper domain in the sense of a conceptually coherent category), there was no terminology specified nor would it be expected that particular terminology would apply, although annotators were instructed to mark terminology if they understood a text to be using domain-specific terminology. In this case the results are expected to differ in subsequent annotation using domain data sets. Similarly, although Locale convention shows reasonable IAA, it had few occurrences and the IAA level shown cannot be considered significant: technical text, however, is more likely to show locale-specific formatting concerns and it is anticipated that this category will prove more important in future annotation rounds.

18 Note that this result contrasts with F-scores, where agreement on the lack of an issue counts towards higher scores. As a result, if two annotators find no instances of an issue, an F-score would indicate 100% agreement while this measure would indicate 0% likelihood of finding the same issue in both. F-score thus indicates overall likelihood of agreement but is not suitable for looking at low-frequency issues. This method addresses only those segments were a particular issue type is found by at least one annotator, and so is useful for looking at low-frequency issues.
A  Annex A: Description of previous versions of MQM

The Multidimensional Quality Metrics (MQM) framework for describing and defining translation quality assessment metrics was developed in the EU-funded QTLaunchPad project. The first, very early, version of MQM was under development when the opportunity arose to have it standardised in the Internationalization Tag Set (ITS) 2.0 specification, where it was realised as the “localization quality issue type” data category (see http://www.w3.org/TR/its20/#lqissue-typevalues). Subsequent development of MQM resulted in a second major version that was released in 2013 and updated throughout 2013 and 2014, with a final update to the pre-harmonisation version made in January 2015.

MQM consists of a number of components. The most crucial is the error hierarchy, which as of the finish of QTLaunchPad consisted of 112 translation quality “issue types” (i.e., descriptors of problems that arise in translation). MQM also has a default method for generating a translation quality score (expressed as a percentage), a mapping to ITS 2.0 and SAE J2450, and an XML representation format. The issue types represent a non-strict superset of existing metrics and error categorizations used in the translation and localization industry.

MQM addresses various sorts of assessment, but has focused primarily on analytic (also known as error-count) metrics, which focus on identifying specific errors (with respect to a given set of translation specifications) within texts. In addition, MQM can be used to define issues in a holistic metric, although the emphasis here is on analytic metrics.

Note, however, that MQM does not assume that there is in fact one perfect, ideal translation, but rather that the ideal is determined with respect to translation specifications that are used to determine what does or does not matter in a translation. Accordingly, MQM does not define a single metric, but rather provides a vocabulary for describing arbitrary metrics. For example, a metric could be created that focuses only on how fluent the translation is and that ignores all other aspects. MQM does not specify what must be checked, but rather provides ways to describe what is checked. (MQM does support holistic evaluation, although this method is not used in QT21 given its focus on diagnostic evaluation.)

At the highest level, MQM issue types were grouped into seven “dimensions”—high-level groupings of issue types. The dimensions were:

- **Accuracy.** Issues related to the relationship of the meaning conveyed by the target text to that of the source, e.g., mistranslations, omissions of content, additions.
- **Fluency.** Issues related to the linguistic well-formedness of content, without respect to whether the content is translated or not, e.g., spelling mistakes, grammar errors (that do not change the meaning), typography errors, stylistic problems.

---

19 Online copies of public versions are maintained at http://www.qt21.eu/mqm-definition/
20 The listing is non-strict in that it excluded some details seen in specific metrics. For example, one popular quality-checking tool identified eight different types of errors with whitespace in text, but MQM contains only one: if additional types are needed (such as the various types of whitespace errors), users can add custom categories, something that was done in the QTLaunchPad project for analysis of MT results.
- **Verity.** Issues related to the content in relation to the world, e.g., content that is legally problematic in one region, content that does not apply to a particular locale. Note: Verity had not been previously treated as a separate aspect of translation quality in translation studies, even though it is a fundamental concern for localization and closely related to the overt/covert distinction in translation studies. It allows evaluators to address cases where text is accurately and fluently translated but where it may not be appropriate for a particular audience or locale. (Some limited aspects of Verity have often been considered under terms such as “Market compliance testing”.)
- **Design.** Issues related to the formatting/layout of content, e.g., font choices, bold face, italics, use of colours.
- **Internationalization.** Issues related to the “translatability” or international engineering of content, e.g., embedded text in images, hard-coded date and time formats. Note: Internationalization had no defined subtypes in previous versions of MQM.
- **Compatibility.** Contains legacy issues from the LISA QA Model that did not relate to translation product quality and so were out of the scope of MQM.\(^1\) Use of these issues is deprecated.
- **Other.** Used for any issues that cannot be categorized into other dimensions.

Dimensions provide descriptions of general areas of content quality. Each dimension can be treated as an issue type itself, but also includes all of the individual descendendent issues types lower in the hierarchy, which can be used when more granularity is needed. For example, a metric could use Accuracy as an issue type by itself, or it could use Accuracy and its individual children nodes, such as Mistranslation, Omission, and Addition. Both approaches are allowable in MQM and the latter can be reduced to the former since the children have a type of relationship to their parent.

The entire MQM hierarchy (as of November 2014) can be visualized as shown in Figure 4:

---

\(^1\) MQM focuses exclusively on issues that relate to the translation itself (i.e., the translation as a product). It does not address the translation process (i.e., the way in which the translation is produced) or ancillary issues such as timeliness of delivery, adherence to security requirements, etc. Nor does it address functional quality of software. These issues were addressed in the LISA QA Model and are important considerations, but beyond the score of MQM.
Figure 4. Pre-harmonisation MQM hierarchy of issue types (November 2014)

It is important to note that children of any issue in the hierarchy are not intended to be exhaustive. For example, the issue type Grammar has three subtypes: Word form, Word order, and Function words. They represent common specific types, but they do not contain all possible types of grammatical errors. For example, the English sentence “He slept the baby” contains what is technically known as a valency error (it treats an intransitive verb as a transitive verb). Since there is no specific type for
valency errors (which are considered highly specific and which are not included in any translation quality assessment metric encountered in QTLaunchPad), this sentence would be categorized as containing a Grammar error (alternatively, MQM can be extended to include new, user-specific, issue types, see below). 22

MQM is flexible and user-extensible, but as a result, it does mean that MQM metrics and their resulting scores may not be directly comparable to one another since they may not check the same issues, unless they can be abstracted to a common subset of MQM. For example, if one metric checks various types of Mistranslation and the other checks only Mistranslation, the results are still comparable; however, if one metric checks issues that do not fall with the children nodes of any issue checked by another, the scores will not be comparable. By taking a maximalist approach MQM has considerable expressive power, but detail may be lost when interchanging with a less granular metric or one that does not address certain aspects of the translation (e.g., if a metric designed for a technical translation is applied to an advertisement it is highly likely that it will not consider all relevant aspects of the advertisement).

---

22 It is highly unlikely that the "official" MQM would ever have such categories added given that few non-linguists would recognize the nature of such an error, although most would recognize the example as a grammatical error.
B  Annex B: Description of DQF

B1.  DQF in General

DQF is a comprehensive suite of tools for quality evaluation of both human and machine translation developed by TAUS. It offers adequacy (in the traditional MT sense), fluency and error-typology based evaluation, productivity measurement, content profiling and a knowledge base containing best practices and use cases. The framework has been developed in close cooperation with many of the TAUS member companies. In 2014 TAUS released the DQF tools running on the TAUS web site.

B2.  Quality Dashboard

Recently TAUS has launched the Quality Dashboard. Users including vendors and buyers of translation services, quality managers, project managers and freelance translators can track and benchmark quality, productivity and efficiency of translation through an open API that connects their translation tools and workflow systems with the Dynamic Quality Framework (DQF) developed over the past four years with and for TAUS members. MateCat, TRADOS Studio and WorldServer are among the translation tools that already connect to the Quality Dashboard. The DQF API now makes it possible for everyone to use DQF from within their own translation tool environment.

The power of data manifests itself also in the new Efficiency Score that is released together with the Quality Dashboard. The Efficiency Score combines throughput per hour with a count of edits, keystrokes and mouse clicks and calculates a weighted score, which gives a much more balanced and realistic insight in the quality and performance of both human and technology resources than the commonly used productivity measurement.

B3.  Error-typology Evaluation

A vast majority of the providers and buyers of translation services manage their quality program with an error typology template. The LISA QA model that was developed in the 1980s forms the basis of the error categories being applied in most cases. The SAE J2450 is another metric that is well-known in automotive translations. TAUS worked with Dr. Sharon O’Brien to develop 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. Content (or a random sample of it) is evaluated by a qualified linguist who flags errors, applies penalties and establishes whether the content meets a pass threshold. This is the type of evaluation that is currently common in the translation sector. Although the error categories might vary, there is broad agreement about the types of errors, but less agreement on the penalties to be applied or their severity levels. This type of evaluation would be recommended for content that was rated high for utility and/or sentiment, but where the time factor was rated lower and would be carried out by a qualified linguist.

B.3.1  Background

"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. It is a largely manual process, focused
only on small samples and takes time and money to apply. The most commonly used categories are: Language, Terminology, Accuracy and Style.

B.3.2 Scenarios

The error-typology approach is used when you are interested in identifying and classifying all errors in the text. Both Utility\(^{23}\) and Sentiment\(^{24}\)-related issues need to be pinpointed by an expert linguist. Project deadlines and volumes are known and time can be allocated to the evaluation. The first hand opinion of customers is not necessary and therefore, they don’t need to be involved in the evaluation process.

Another scenario where this model can be used is one in which the translation quality provided by the vendor is reviewed after translations have been released. Also, it can be used at a later stage to identify specific mistakes made by translators or the Machine Translation system and guide their improvement.

B.3.3 Overview

This evaluation model involves the use of a translation error typology. Content quality is judged based on the quantity and severity of the errors found. Content (or a random sample of it) is evaluated by a qualified linguist who flags errors, applies penalties and establishes whether the content meets a passing threshold. This type of evaluation is currently commonly used in the translation industry.

This model can be used to identify errors in text translated automatically and by human translators. The error category can be adapted to suit different MT types and (computer-aided) human translation processes (see the Articles section).

B.3.4 Evaluation Approach

The content, or a random sample of it, is provided to a qualified linguist. The linguist is given an error typology and asked to carefully consider the source and target segments and to highlight any errors.

Error typologies vary, but a recent benchmarking report by TAUS found that there was considerable similarity between the typologies being used by over 20 companies. The most commonly used categories for User Assistance content are Language, Terminology, Accuracy and Style. Diagnostic evaluations which seek to understand in detail the nature or cause of errors may require a more detailed error typology. The typology needs to be flexible enough to allow for additional or sub-categories, if required.

The errors typically have four severity levels: critical, major and minor and neutral. “Neutral” applies when a problem needs to be tagged, but is not the fault of the translator.

Penalties are associated with each error and severity level.

The value of a pass/fail threshold is determined by the types or errors, their frequency, and the severity level attributed to them. The thresholds are flexible and dependable on content type, end-user profile and perishability of the content.

---

\(^{23}\) Utility refers to the relative importance of the functionality of translated content.

\(^{24}\) Sentiment refers 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.
Once a penalty threshold is exceeded, the translation is deemed to have failed the quality evaluation. Different thresholds might hold for different content types or even target languages.

**B4. Error-typology in DQF tools**

Error typology-based evaluation is available in the DQF tools. It enables you to categorize and count translation errors segment-by-segment using commonly used industry criteria for – accuracy, language, terminology, style and country standards.

https://www.taus.net/knowledgebase/images/a/a6/Taus_dqf_errortypology_evaluation_template.xlsx Download the extended DQF error-typology template

**B5. Definitions for the main error categories**

<table>
<thead>
<tr>
<th>Category</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language</td>
<td>Although it can refer to ambiguous sentences, an error in this category generally means a grammatical, syntactic or punctuation error.</td>
</tr>
<tr>
<td>Terminology</td>
<td>A glossary or other standard terminology source has not been adhered to.</td>
</tr>
<tr>
<td>Accuracy</td>
<td>Incorrect meaning has been transferred or there has been an unacceptable omission or addition in the translated text.</td>
</tr>
<tr>
<td>Style</td>
<td>Quite subjective, it refers to a contravention of the style guide.</td>
</tr>
</tbody>
</table>

**B6. Definitions for the severity levels**

<table>
<thead>
<tr>
<th>Severity</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Severity 1</td>
<td>Critical Errors that may carry health, safety, legal or financial implications, violate geopolitical usage guidelines, damage the company’s reputation, cause the application to crash or negatively modify/misrepresent the functionality of a product or service, or which could be seen as offensive.</td>
</tr>
<tr>
<td>Severity 2</td>
<td>Major Errors that may confuse or mislead the user or hinder proper use of the product/service due to significant change in meaning or because errors appear in a visible or important part of the content.</td>
</tr>
<tr>
<td>Severity 3</td>
<td>Minor Errors that don’t lead to loss of meaning and wouldn’t confuse or mislead the user but would be noticed, would decrease stylistic quality, fluency or clarity, or would make the content less appealing.</td>
</tr>
<tr>
<td>Severity 4</td>
<td>Neutral Used to log additional information, problems or changes to be made that don’t count as errors, e.g. they reflect a reviewer’s choice or preferred style, they are repeated errors or instruction/glossary changes not yet</td>
</tr>
<tr>
<td>Implemented, a change to be made that the translator is not aware of.</td>
<td></td>
</tr>
<tr>
<td>---------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td><strong>Kudos</strong></td>
<td>Used to praise for exceptional achievement.</td>
</tr>
</tbody>
</table>
Figure 5. DQF Error Typology (2014)
C  Annex C: Listing of issues types in DQF subset

The following listing is taken from the online MQM definition at http://www.qt21.eu/mqm-definition/definition-2015-06-16.html#dqf-mapping:

- **Accuracy** (accuracy)
  - Addition (addition)
  - Improper exact TM match (improper-exact-tm-match)
  - Mistranslation (mistranslation)
    - Ambiguous translation (ambiguous-translation)
    - Mistranslation of technical relationship (technical-relationship)
    - Overly literal (overly-literal)
  - Omission (omission)
    - Omitted variable (omitted-variable)
  - Over-translation (over-translation)
  - Under-translation (under-translation)
  - Untranslated (untranslated)

- **Design** (design)
  - Length (length)
  - Local formatting (local-formatting)
  - Markup (markup)
  - Missing text (missing-text)
  - Truncation/text expansion (truncation-text-expansion)

- **Fluency** (fluency)
  - Character encoding (character-encoding)
  - Grammatical register (grammatical-register)
  - Grammar (grammar)
  - Inconsistency (inconsistency)
    - Inconsistent with external reference (external-inconsistency)
  - Link/cross-reference (broken-link)
  - Punctuation (punctuation)
  - Spelling (spelling)

- **Locale convention** (locale-convention)
  - Address format (address-format)
    - Postal code (postal-code)
  - Date format (date-format)
  - Currency format (currency-format)
  - Measurement format (measurement-format)
  - Shortcut key (shortcut-key)
  - Telephone format (telephone-format)

- **Style** (style)
  - Awkward (awkward)
o Company style (company-style)
o Inconsistent style (inconsistent-style)
o Third-party style (third-party-style)
o Unidiomatic (unidiomatic)

• Terminology (terminology)
  o Inconsistent with termbase (termbase)
    ▪ Company terminology (terminology-company)
    ▪ Third-party termbase (terminology-third-party)
  o Inconsistent use of terminology (term-inconsistency)

• Verity (verity)
  o Culture-specific reference (culture-specific)