Multidimensional Quality Metrics: A Flexible System for Assessing Translation Quality

Arle Richard Lommel
DFKI Berlin
arle.lommel@dfki.de
Alt-Moabit 91c
D-10559 Berlin, Germany

Aljoscha Burchardt
DFKI Berlin
aljoscha.burchardt@dfki.de
Alt-Moabit 91c
D-10559 Berlin, Germany

Hans Uszkoreit
DFKI Berlin
hans.uszkoreit@dfki.de
Alt-Moabit 91c
D-10559 Berlin, Germany

Arle Lommel is a Senior Consultant at the Berlin project office of the German Research Center for Artificial intelligence (DFKI)'s Language Technology lab. Formerly employed by the Localization Industry Standards Association (LISA) and the Globalization and Localization Association (GALA), he is an expert in translation quality assessment-related topics.

Aljoscha Burchardt is a Senior Researcher at the Language Technology Lab of the German Research Center for Artificial Intelligence (DFKI). He manages several projects dealing with different aspects of Machine Translation such as hybrid translation (taraXŰ) or measuring translation quality (QTLaunchPad).

Hans Uszkoreit is Scientific Director and Head of the Language Technology Lab at DFKI, also Professor of Computational Linguistics and Computer Science at Saarland University since 1988. He has held positions at Stanford, SRI, and IBM Germany. He is a Member of the European Academy of Sciences, the Intl. Committee for Computational Linguistics, the ELRA Board, and of several advisory and editorial boards. He has coordinated several large EU and national projects, including EuroMatrix, EuroMatrix Plus, and QTLaunchPad on new approaches to machine translations, and META-NET.

1. Introduction

In the last decades translation quality has emerged as a major issue in the translation activities international businesses. Unfortunately, "quality" itself has been an elastic concept that often amounts to highly variable impressions from individuals. Specifications such as SAE J2450 and the LISA QA Model have attempted to ameliorate this situation, but are one-size-fits-all models that cannot be easily adapted to fit various needs. In addition, methods for assessing human and machine translation have been totally disconnected, rendering meaningful comparison between them impossible. An additional problem that arises is that issues in source texts often remain undetected until translation time, at which point translators have to make sense of defective texts and are then blamed for problems that arise because of problems in the source.

In response, the EU-funded QTLaunchPad project has developed the "multidimensional quality metrics" (MQM) framework for assessing quality for both translated texts and source texts in the context of translation. This framework is based on an analysis of over twenty existing translation quality assessment metrics and provides a large set of issue types that may be checked in both human and machine translation. It provides a list of over 100 issue types that cover all of the major translation quality assessment metrics. These issue types are intended to serve as a "master catalog" from which relevant issues for assessing the quality of specific types of translation quality tasks may be chosen. MQM is free and open and can thus be implemented and extended to match any needs.

This paper describes the MQM definition of quality and the MQM system, issue types, and methods. It also gives pointers to existing software and tools.

2. Defining quality

The definition of translation quality has long been an issue in academic translation studies. Much of the traditional focus in (human) translation studies has been derived from literary translation practice assuming the existence of absolute quality and the availability of unlimited resources. In the 1990s, however, software localization began to emerge as a separate branch within translation and its business-oriented requirements tended to focus on identify and quantifying individual errors to produce a quality score that could be used for acceptance testing and to support other business decisions. Developments in localization led to the creation of many company-specific metrics for assessing translation quality, but also to moves to standardize quality processes. Two relatively successful metrics were the LISA QA Model (initially developed in the 1990s and last updated in 2006 (LISA 2006))—intended for software and document localization—and SAE J2450 (released in 2001), which addressed the needs of automotive service manuals (SAE International 2001). In the absence of other accepted models, both of these metrics were widely implemented, even for tasks beyond their original scope of application.

However, the models developed in the 1990s and 2000s were, despite disclaimers about applicability, generally treated as one-size-fits-all models used for all types of translation tasks. They thus assume, intentionally or not, the "transcendent" perspective on translation quality, which would maintain that translations are either right or wrong when compared to an ideal standard. In this view quality is a product of the target text in relationship to the source text, regardless of use.

The ever-increasing pressure on the translation industry (in terms of price, volume, and turnaround time) and the emergence of commercial applications for Machine Translation (MT), particularly for "just in time" or "on demand" applications, have challenged the notion of absolute translation quality. MT applications often supply translation in cases where issues of timeliness or accessibility trump abstract notions as to whether a translation is "correct" or not. In these instances a translation may be deemed "good enough" for a purpose. In addition, MT assessment has largely been based on the mechanical similarity of MT output to human reference translations, an approach exemplified in quality measures like BLEU and NIST.

The notion of purpose has also become more important as it has become apparent that, for example, the requirements for translating a online service document (where the translation must serve the functional purpose of helping a reader solve a particular problem) are different from the requirements of translating a regulatory notice (which compliance with legal requirements and preventing harm are the highest priorities). Similarly, both of these have very different requirements than the requirements need to translate subtitles used in a popular media program.

It is important to note that discussions about translation quality (for both human and machine translation) have been conducted in a vacuum, largely independent of broader business theory about quality. This broader business literature (here exemplified by Garvin 1984) has focused on multiple definitions of quality and a functionalist perspective that defines quality in terms of whether products, processes, and projects conform to expectations. In such a view, a product is deemed to exhibit quality if it meets all stated requirements, rather than by being judged compared to an abstract notion of quality. In particular Garvin emphasized the importance of *manufacturing quality* (i.e., whether all specifications are met) and *user quality* (i.e., whether the product meets the requirements of users). These perspectives have been applied successfully to many fields, but until recently were not applied to translation theory.

Based on the broader literature on quality, Alan Melby, working with one of the authors (Lommel), developed a proposed universal definition of "quality" for translation:

> A quality translation demonstrates required accuracy and fluency for the audience and purpose and complies with all other negotiated specifications, taking into account end-user needs. (Melby forthcoming, Lommel 2013)

Although this definition is quite simple in its formulation, it makes a sharp break from traditional translation quality definitions in favor of a Functionalist, Skopos-oriented perspective on translation quality. It attempts to unify the transcendent, manufacturing, and user perspectives on quality. While absolute transcendence is rejected, it does recognize that accuracy and fluency can be conceived of in absolute terms and that the default expectation in many translation contexts is for full accuracy and fluency. (Here "accuracy" is understood in terms of how well the target text represents the informational content of the source text and "fluency" is understood as referring to features such as grammaticality, clarity, and format that apply to any text, regardless of its status as a translation.)

Taking this definition as a basis, translation quality can only be assessed in terms of whether or not a translation meets specified requirements and meets its communicative purpose. Since requirements for accuracy and fluency vary, specifications are not always identical, and different groups of end-users have different requirements, it necessarily follows that no single, fixed metric can be used to assess all translation quality. Such a monolithic model of quality would return assessments inappropriate for specific tasks and would end up holding the translation of a literary work and a weather bulletin to the same standard. Instead a flexible and adaptable framework that can account for specific needs is required.

3. Multidimensional quality

The Multidimensional Quality Metrics (MQM) system for developing translation quality assessment metrics, developed as part of the EU-funded QTLaunchPad (*http://www.qt21.eu/launchpad*) project, operationalizes the definition of translation quality described in the previous section. It provides a "catalog" of issue types (108 types as of November 2013) arranged in a hierarchy (see ***Figure 1***) (QTLaunchPad 2013a). The list of issues included in the hierarchy was created by a rigorous comparison of the issue types identified by major quality metrics (LISA QA Model, version 3.1, SAE J2450, ISO CD 14080 (cancelled), SDL TMS Classic Model, American Translators Association Certification Grading Criteria, TAUS Dynamic Quality Framework) and quality checking tools (Acrocheck, ApSIC XBench, CheckMate Quality Check, LanguageTool, QA Distiller, XLIFF:doc) as well as a comprehensive

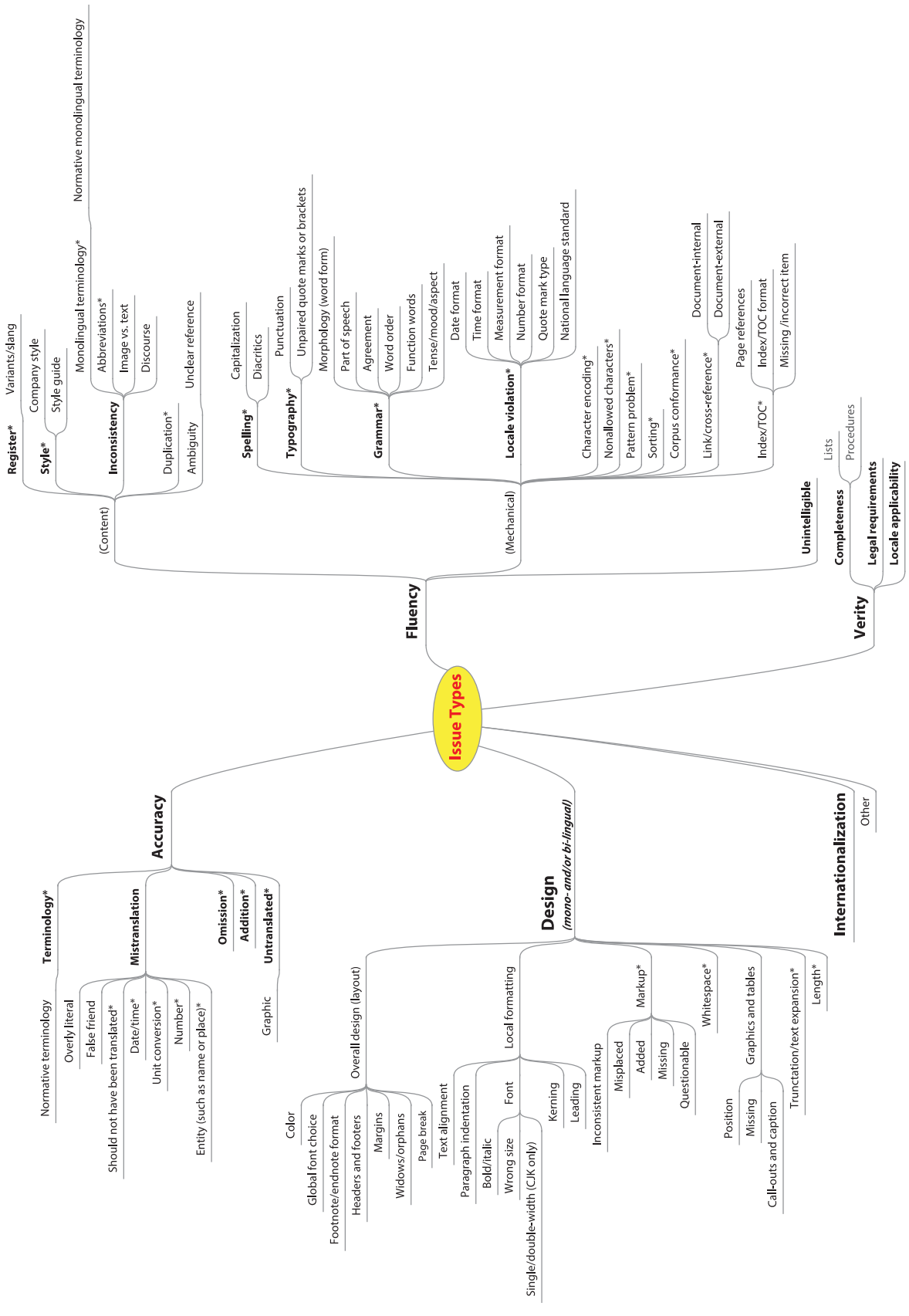

Figure 1. *Full MQM hierarchy. Issues with an asterisk (*) by them are ones where automatic detection of issues may be possible.*

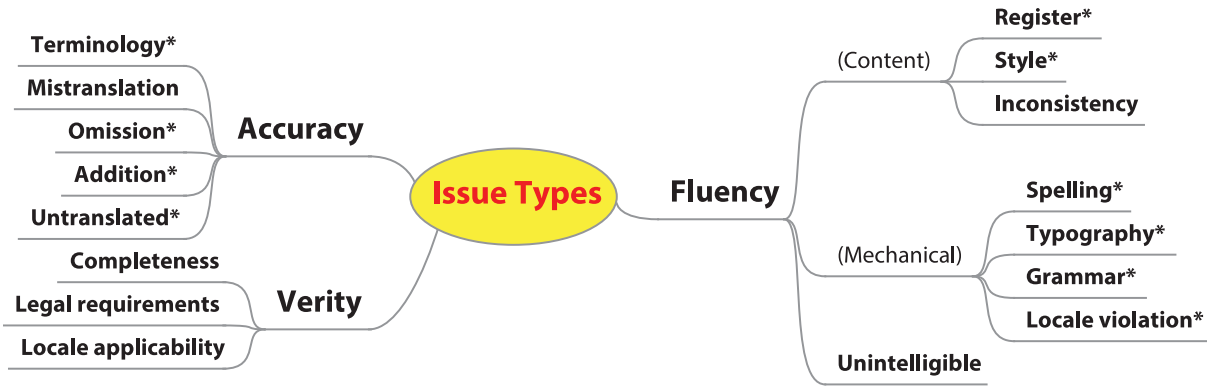

Figure 2. *MQM Core.*

list of major MT quality measures and tools (Adequacy and Fluency, Rankings, Error Analysis, BLEU, NIST, METEOR, WER/PER/(h)TER/TERp, Evaluating Post-Editing Effort, Quality Estimation, MT Error Classification and Diagnostic MT Evaluation, Tools for MT Error Analysis, Addicter, AMEANA, BLAST, DELiC4MT, Hjerson, TerrorCat, Woodpecker, Hjerson error categories). The categories listed represent an attempt to create a non-strict superset of the concrete issue types identified in the measures and tools listed above. Since the MT evaluation metrics and tools tend not to identify specific types of issues, they contributed very little to the listing of issue types. Because the various metrics considered differed in their granularity, some particularly granular distinctions (such as the distinction between eight individual types of whitespace errors in the Okapi Checkmate system) were omitted.

The resulting list supports multiple levels of granularity and abstraction. As shown in ***Figure 1***, the hierarchy presents various items as subtypes of higher-level categories. For example, *Morphology (word form)* is considered a subtype of *Grammar*, which is in turn considered a subtype in them *Mechanical* branch of Fluency. The children of any particular node are considered exemplars of their parent, but are not intended to be an exhaustive enumeration of all types of the parent. As a result a quality check might identify issues within *Grammar* but find issues that could not be given a precise categorization within *Grammar*, which should then be assigned to *Grammar.*

(Note that MQM can be interpreted as a common vocabulary for describing metrics. Concrete existing metrics can easily be linked to it as it is free and open. The QTLaunchPad project currently is in the process of handing MQM over to the community—see below—so that it can be maintained and extended in the future.)

MQM categorizes issues into five "branches":

- **Fluency**. Issues related to the language of the translation, regardless of its status as a translation.
- **Accuracy**. Issues related to how well the content of the target text represents the content of the source.
- **Verity**. Issues related to how the text corresponds to real world requirements.
- **Design**. Issues related to the formatting and layout of the text.
- **Internationalization**. Issues related to the internationalization (engineering for localizability) of the content. (Note that this branch is currently underspecified but is intended for future development.)

The *Verity* branch requires some explanation as it represents issues not typically treated as part of translation assessment. These issues have to do with the conformance of the text to the external world of the text. For example, if a German translation of an electrical device manual accurately translates a statement that says that a ground wire is bare copper—true in the United States—it will be incorrect in Germany, where it would be covered in green and yellow striped insulation. In such situations there is a problem with the text, even if it accurately translates the source. The inclusion of issues related to verity in MQM marks a novel contribution to the discussion of translation quality assessment in a business setting, taking some burden off the translators who are often made responsible for any type of deficit in a translation irrespective of whether it is their fault or not.

Notes that items in the Fluency and Accuracy branches can also be applied to source texts to verify their quality, and MQM provides a scoring mechanism that allows for translators to be credited for improvements they make over the source.

It is also important to note that it is not expected that any task would utilize all (or even most) of the categories in the full MQM set. Instead quality assessment tasks are intended to make a relevant selection of issues from the full list. In most cases we anticipate that users would select no more than twelve issue types, although more detailed analytical needs might require more. To simplify things, MQM features a "core" of 19 issues appropriate for many general-purpose assessments of plain text translation (***Figure 2***). The recommendation is that issues be selected from the core unless a task requires finer granularity or use of issues not included in the main branches available in the Core. Additional items not found in the core, such as issues related to formatting or more detailed issue types,

are found in MQM extensions that can be used as needed. However, by limiting choices to the core where possible, greater interoperability in quality metrics can be fostered and—hopefully—helps different user groups to define their own useful subsets.

Even the Core may be too detailed for some tasks. Since a large listing of issue types alone can be counterproductive, MQM provides a way to use 12 "dimensions" (based on the ISO/TS 11669 (ISO 2012) specification for translation specifications) to help guide users to select those issues that relate to project requirements and expectations. The dimensions are as follows:

1. (Target) Language/locale
2. Subject field/domain
3. Terminology (source/target)
4. Text type
5. Audience
6. Purpose
7. Register
8. Target text style
9. Content correspondence
10. Output modality*
11. File format
12. Production technology

While a full discussion of these dimensions is beyond the scope of the paper (see *http://www.ttt.org/specs* for a more detailed explanation of most of the dimensions), a few notes are required for dimensions whose meaning may not be immediately apparent:

- *Output modality* (dimension 10) deals with the way in which text will be presented, regardless of the file format. For example, text presented on a mobile device's screen or via an embedded display in an industrial manufacturing device may have strict length limitations, or subtitles may require that information be cut from discourse in order to be readable. Knowing the intended mode of output may influence translation quality expectations.
- *Content correspondence* (dimension 9) is a broad dimension that deals with how the source and target text correspond. For example, a translation may be *overt* (it does not hide the fact that it is a translation, perhaps preserving textual features specific to the source language) or *covert* (it attempts to appear as though it were written in the target language and hides the fact that it is a translation); a translation may be a complete translation, a summary translation, or a gist translation; and so forth.

These dimensions answer basic questions about the text, its intended audience and usage, and the translation process. By selecting only the set of task-relevant issues, users can be assured that a quality score addresses their needs. Since these dimensions relate to the standard translation specifications created using ISO/TS 11669, assessment can be tied into procurement: from the moment a job is put out for bid, the assessment criteria are already being developed. As a side effect, a formal specification during negotiation with the customer of requirements concerning translation and quality assessment helps to minimize misunderstandings caused by "default assumptions" and to precisely pinpoint who is responsible if a translation does not meet expectations.

The process of developing an MQM-compliant quality assessment metric involves identifying those issues that would enable a reviewer to verify that a translation meets the requirements set forth in the dimensions. For example, if a translation is intended to be a full translation, *Omissions* would generally be marked as errors, but in the case of a translation intended to be displayed as subtitles, omission is to be expected and omissions should either be ignored or noted only when they cause a major problem in understanding the video for which they are translated. Similarly a metric for assessing service manuals (a text type) for service technicians (an audience) might omit any consideration of *Style* since the text does not need to exhibit good style to be useful, while an advertising text intended to persuade a general audience might put a very high priority on *Style*.

In order to assist users in building metrics, the QTLaunchPad project has created a set of demonstration tools (available at http://www.qt21.eu/MQM/) that guide users through the process of building a set of dimensions, selecting issue types to create a relevant metric, and using this metric to assess translations. While these tools are relatively simple (and support only the MQM Core at present), more sophisticated tools (e.g., featuring drag-and-drop functionality for issue selection and improved ergonomics) are possible.

Note as well that the comprehensive nature of the full MQM list is intended to allow existing metrics to be mapped easily to MQM so that they can be ported with minimal loss to MQM. As a result, if assessors have a met-

ric that meets their needs, there is no obligation to change the metric, but it can be described in MQM to facilitate comparison with other metrics and portability.

4. Scoring

Because the ability to generate scores is important in many environments, MQM provides a scoring system, defined as described below. The following basic formula is used for calculating MQM quality scores in an error-count environment:

$$TQ = 100 - AP - (FP_T - FP_S) - (VP_T - VP_S)$$

Where:

- TQ = quality score. *The overall rating of quality*
- AP = penalties for Accuracy. *Sum of all weighted penalty points assigned in the Accuracy branch*
- FP_T = Fluency penalties for the target. *Sum of all weighted penalty points in the target text assigned to the Fluency branch. (Note: for computational purposes, Design and Internationalization are treated with Fluency.)*
- FP_S = Fluency penalties for the source. *Sum of all weighted penalty points in the source text assigned to the Fluency branch. If the source is not assessed $FP_S = 0$.*
- VP_T = Verity penalties for the target. *Sum of all weighted penalty points in the target text assigned to the Verity branch.*
- VP_S = Verity penalties for the source. *Sum of all weighted penalty points in the source text assigned to the Verity branch. If the source is not assessed $VP_S = 0$.*

Note that because penalties for source Fluency and Verity are added to the score, it is possible for a translation to have an overall score greater than 100 if the target-language text represents an improvement over the source. Thus this formula, when fully applied, recognizes translators for work that fixes problems found in the source.

Penalties are counted according to the following default formula, which takes default severity multiplier values from the LISA QA Model. It defines minor issues as those that do not impact meaning or usability; major issues as those that impact meaning or usability but do not render the text unusable; and critical issues as those that render a portion of the text unfit for purpose.

$$P = (Issues_{minor} + Issues_{major} \times 5 + Issues_{critical} \times 10) \div Word\ count$$

These multipliers can be customized and alternative severity levels used (for example, many metrics use only two severity levels). *P* (penalty) values can be calculated for any single issue or group of issues. All penalty values are summed up within a branch. For example, if a metric checked only Terminology and Mistranslation in the Accuracy branch and penalties for Terminology = 1.2% and penalties for Mistranslation = 1.4%, then PA_T would equal 2.6%. One unresolved issue in MQM is to determine whether these particular severity multipliers are valid in general.

A scorecard (the so-called "Tabular Scorecard") that displays all of the math and assumptions (and allows issue weighting and severity levels to be customized) is available in the MQM online resources at *http://www.qt21.eu/MQM*. Where customization and display of all intermediate calculations is not required, a "Simple Scorecard" with a streamlined interface is also available.

5. MQM Tools, implementation, and future plans

As discussed above, the QTLaunchPad project has provided tools to demonstrate the use of MQM. These tools assist users in defining dimensions, building appropriate metrics, and implementing them. The source code for these tools will all be made available under the Eclipse Public License in early 2014. They are intended to provide a technology demonstration and are not intended for production use. Further tools development may be conducted in the future. The MQM framework and software is released under an open license that will allow maximum reuse. Any parties interested in implementing MQM—in free or commercial software—will be able to do so free of charge.

In addition to these tools, the QTLaunchPad project has funded development of an open- source translation-editing environment, translate5 (http://www.translate5.net), that allows users to mark issues in line with MQM-compatible markup. The project is currently engaged in improvements in ergonomics and reporting that will make it more user-friendly. MQM source code is currently available under the GPL 3.0 license with FLOSS exceptions (to make the licensing terms less restrictive) to promote further development and implementation in open-source

projects. Translate5 is currently in use in production by a number of language service providers (LSPs), so it is anticipated that inclusion of MQM in it will help foster adoption.

Because funding for the QTLaunchPad project runs out in 2014, the project is currently in the process of handing over the MQM specification to the CRISP program (http://www.gala-global.org/goals-overview) of the Globalization and Localization Association (GALA) for long-term development and maintenance by the localization industry. CRISP was chosen for this action because it does not require membership fees and is designed to be open and accessible to interested parties. It is anticipated that MQM will be further refined in close association with industry input. In addition, preliminary discussion is underway to develop formalisms for representing MQM in XML and HTML5 within the World Wide Web Consortium's Internationalization Interest Group.

When work has matured sufficiently in both the definition of MQM issue types and the formalisms for representing MQM metrics and quality data, the intention is to submit MQM for further development by a recognized standards body, such as OASIS, and eventually to ISO with the goal of achieving international standard status for MQM.

6. Conclusion

Preliminary and ongoing testing indicates that MQM is successful in describing a wide variety of quality metrics within an error-count environment. Metrics represented in MQM have been successfully used for assessing the quality of both human and machine translation. In the case of machine translation results, investigation found that MQM delivered quality assessment results that accorded well with results obtained from post-editing tasks (QTLaunchPad 2013b), but added analytic insight beyond what is available post-editing and post-editing-based automatic error classification. Because MQM is available under open terms and has a long-term plan for viability and implementation, we are optimistic that it will help address some of the long-standing problems with translation quality assessment.

7. Acknowledgements

The QTLaunchPad project is funded by the 7th Framework Programme of the European Commission through the contract 296347.

8. References

Garvin, David. 1984. "What does 'product quality' really mean?" *Sloan Management Review*, fall, pp. 25–45.

ISO. 2012. *ISO/TS 11669:2012: Translation projects -- General guidance*. Geneva: ISO.

LISA. 2006. "LISA QA Model 3.1: Assisting the localization development, production and quality control processes for global product distribution" (press release). Romainmôtier: LISA.

Lommel, Arle R. 2013. "Measuring translation quality in today's automated lifecycle". Presentation at tcworld 2013, Wiesbaden, Germany, November 8. Available from *http://tagungen.tekom.de/fileadmin/tx_doccon/slides/453_A_Unified_Model_for_Document_and_Translation_Quality_Assurance.pdf* (accessed November 14, 2013).

Melby, Alan K. forthcoming. "Human and machine translation quality: Definable? Achievable? Desirable?". To appear in *lacus forum* 39.

QTLaunchPad. 2013a. *Multidimensional Quality Metric quality issue types*, version 2.5.5. *https://docs.google.com/document/d/1E8IR1-8bR_M7VouHQhogUPpP2htpprwMMx1Mk9KPBTI/pub* (accessed November 14, 2013).

———. 2013b. *Error analysis in near miss translations*. http://www.qt21.eu/launchpad/deliverable/error-analysis-near-miss-translations (accessed November 14, 2013).

SAE International. 2001. *SAE J2450: Translation Quality Metric.*