2015-06-01

Establishing the Viability of the Multidimensional Quality Metrics Framework

Tyler A. Snow
Brigham Young University - Provo

Establishing the Viability of the Multidimensional Quality Metrics Framework

Tyler A. Snow

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Arts

Alan K. Melby, Chair
Deryle Lonsdale
Daryl Hague

Department of Linguistics and English Language

Brigham Young University

June 2015

ABSTRACT

Establishing the Viability of the Multidimensional Quality Metrics Framework

Tyler A. Snow
Department of Linguistics and English Language, BYU
Master of Arts

The Multidimensional Quality Metrics (MQM) framework is a new system for creating customized translation quality assessment and evaluation metrics designed to fit specific translation needs. In this study I test the viability of MQM to determine whether the framework in its current state is ready for implementation as a quality assessment framework in the translation industry. Other contributions from this study include: (1) online software for designing and using metrics based on the MQM framework; (2) a survey of the typical, real-world quality assessment and evaluation practices of language service providers in the translation industry; and (3) a measurement scale for determining the viability of translation quality assessment and evaluation frameworks such as MQM. The study demonstrates that the MQM framework is a viable solution when it comes to the validity and practicality of creating translation quality metrics for the translation industry. It is not clear whether those metrics can be used reliably without extensive training of qualified assessors on the use of MQM metrics.

Keywords: Translation, Quality, Assessment, MQM, Metric, Reliability, Validity, Practicality, Viability

ACKNOWLEDGMENTS

There are many people who deserve recognition for their contributions. This includes friends, family, and many colleagues around the globe including:

- Alan K. Melby: For his extensive contributions to this study in addition to his many contributions to the entire translation community.
- Deryle Lonsdale and Daryl Hague: For their attention to detail and availability throughout the writing process.
- Arle Lommel: For his knowledgeable expertise in all things MQM.
- Serge Gladkoff: For his collaboration and assistance during the GALA follow-up survey interview process.
- Jeff Beatty: For his hard work and dedication in implementing MQM at the Mozilla corporation.
- Paul Fields: For his expertise in statistics and experimental design.
- My wife, Whitney: For putting up with the many long hours it took to write this thesis paper.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

Multidimensional Quality Metrics (MQM) is a generalized framework for assessing the quality of any translation through the creation of customized metrics tailored to the particular translation requirements. MQM was designed by the QTLaunchPad project, headed by the German Research Center for Artificial Intelligence (DFKI) to provide a flexible method for describing translation quality metrics appropriate for both human and machine translation, allowing them to be compared on an equal basis for suitability according to various specifications. The framework includes over a hundred distinct error categories which are the building blocks for creating customized translation quality assessment metrics. In this study I examined the inter-rater reliability between assessors using one instance of a MQM metric designed to assess the translation quality of the Mozilla Firefox web-browser from English to Mexican-Spanish. In addition, I examined the validity and practicality of that same MQM-style metric. To complete the study, I created a web application that can implement translation quality metrics based on the MQM definition.

In addition to the Mozilla case study, I determined whether MQM contains a complete set of error categories for the assessment of industry translations by interviewing 32 different translation companies about their quality assessment practices.

Through the case study I discovered that the particular assessors in the experiment using the Mozilla Firefox MQM metric do not display a high degree of inter-rater reliability. The assessors were considerably less reliable when it came to *identifying* translation errors but much more reliable at *classifying* errors using MQM.

The interviews revealed that the MQM framework contains all but one of the error categories needed for quality assessment in the translation industry meaning that the framework has a valid set of error categories for that sector[1]. Finally, I argue that a practicality study would not reveal anything useful in the particular case of MQM due to its ability to create optimized metrics for any assessment situation. An impractical metric only indicates that the wrong metric was used. Overall, I find that MQM is a valid solution that can produce practical metrics for quality assessment for the translation industry but that a substantial amount of training is required to reliably utilize MQM-style metrics. Assessors must be well qualified before training on MQM-style metrics. How to best implement that training is still speculative at this point.

[1] As a result of this study, offensiveness has now been added as an error category.

BACKGROUND

A. General

When it comes to translation, "theorists and professionals overwhelmingly agree there is no single objective way to measure quality. Yet every day, translators, editors, revisers, clients and many others nonetheless have to do just this" (Drugan, 2013). One of the contributing factors to this confusion is the lack of consensus on the definition of the term "translation quality assessment."[2] How one defines this term greatly affects the manner by which assessment is conducted. The definition of translation quality assessment was the focus of a three part debate published by the journal *Tradumàtica in December 2014*.

The first part of the series detailed two opposing viewpoints on the definition of the term "translation." The first is the narrow-definition of translation which is:

> Translation transfers a written source text into a written target text of roughly equivalent length. Such a translation conveys all the source text's meaning, making only those adjustments necessary for cultural appropriateness without adding, omitting, condensing, or adapting anything else (Melby et al, 2014a).

Translation in this definition is limited to the relationship between the source and target texts. Any other language related tasks, such as localization[3], are considered additional activates to translation.

---

[2] along with many other terms in the realm of translation, see the glossary of terms section for my working definitions the purposes of this study

[3] the adaptation of content to fit a specific locale

In contrast to the narrow-definition is the broad-definition which is stated here:

> Translation is the creation of target content that corresponds to source content according to agreed-upon specifications. (Melby et al, 2014a).

Anything that goes into producing the target content, such as localization, is included as part of the translation process in the broad definition.

Comparing the two definitions, if one were to use the narrow definition to assess the final localized translation, two separate assessments would be needed, one for the translation and the other for the localization. The distinction between which part of the target text is localization and which part is translation is not precisely clear. By contrast in the broad definition, since any process[4] used to create the target text is considered translation, only one assessment is needed.

In part two of the *Tradumàtica* debate (Melby et al, 2014b) the topic of "quality" is described with emphasis on Garvin's Comprehensive Framework: Five Approaches to Quality (1984). Garvin's five approaches to quality are listed as follows:

1) Transcendent approach: quality of a product or service as an innate characteristic that is both absolute and universally recognizable.

2) Product-based approach: quality is quantifiable based on certain ingredients or attributes.

3) User-based approach: quality is in the eye of the beholder and is determined by end-user needs and biases.

[4] including professional human translation and machine translation

4) Production-based approach: quality is the degree of compliance to pre-determined specifications.

5) Value-based: quality is the amount of benefit of the product over the cost.

The definition of quality in each approach affects the outcome of an assessment in measuring if quality has been achieved.

The third part of the *Tradumàtica* debate (Melby et al, 2014c) focuses on the narrow and broad definitions of translation in regards to quality, essentially combining the topics discussed in part one and two of the series. In the narrow-definition of translation, quality is limited in scope to the fluency[5] of the source and target texts and the accuracy[6] between them. Quality requirements are universal and do not change from translation to translation in the narrow-definition. This means that all translations can be held to the same standard of quality and be assessed using the same method similar to Garvin's transcendent approach to quality. As mentioned previously, no single method of translation quality assessment has yet emerged, indicating that a single standard of quality has not or cannot be determined universally for all translations collectively.

[5] the quality of the source or target as a text on its own. (i.e, is the text linguistically well-formed and understandable?)

[6] the degree of correspondence of meaning interpreted from the source and target texts (i.e., does the reader of the target text understand the same content as a reader of the source text would?)

The broad-definition recognizes that each translation is different and that quality is relative to the unique requirements of a particular translation. This most closely aligns with the production-based approach, as explained by Garvin; however, all five approaches could be utilized in the broad-definition. For example, if end-user needs were pre-determined and used as a basis for determining the quality of a translation, both the production-based and user-based approaches would be in use simultaneously. Pre-determining the product benefit or value and quantifiable attributes allows the broad-definition to encompass even the product-based and value-based approaches as well. Even the transcendent approach can be included in the broad-definition, if we can identify translation quality requirements that are universal for all translations.

Working with the broad-definition covers potentially all five of Garvin's approaches to quality but it comes at a price. Since all translations have unique requirements, translation quality assessments must be tailored to measure the adherence to those distinct requirements. The Multidimensional Quality Metrics (MQM) framework has been designed specifically with the broad-definition in mind. It operates under the definition:

> A quality translation demonstrates accuracy and fluency required for the audience and purpose and complies with all other specifications negotiated between the requester and provider, taking into account end-user needs. (Melby, 2012)

As in the broad-definition accuracy and fluency in an absolute sense cannot indicate the quality of a translation. Both must be understood relative to a set of

requirements based on the purpose of the translation, determined and understood by both the requestor[7] and the provider[8], while taking into account the end-user needs of a particular translation. For example, the fluency requirements for an in-house service manual are likely to be much lower than for a customer-facing marketing piece. Similarly, a service manual will put a high value on accuracy, but an evocative marketing piece may allow the translator considerable liberty to adapt the text for the target audience. The relationship of accuracy and fluency to end-user needs means that when it comes to translation quality, there can be no one-size-fits-all approach since end-user needs vary greatly from translation to translation. What is a perfectly acceptable translation in one context might be unacceptable in another.

MQM was designed as a comprehensive translation quality framework designed to assess any translation through the creation of specific metrics tailored to the needs of each individual assessment scenario. The framework consists of over a hundred clearly defined error categories[9] organized into a hierarchy of dimensions[10] and also includes methods for using those error categories in the form of a metric[11]. This is done by first

[7] those who initiate the translation process

[8] those who provide translation services

[9] a type of error

[10] a family of related error categories.

[11] consisting of dimensions, error categories, weights, and thresholds, to be implemented in an appropriate workflow with accompanying assessment tools

determining the external requirements of a translation which are elucidated by a set of translation blueprints, known as translation *specifications*.[12]

Specifications help explain the requirements and help define what is unique about a particular translation. These specifications are derived from a generalized set of translation *parameters*[13] based on the ISO/TS-11669[14] and ASTM F2575[15] standards which are essentially the same set of parameters. An example of a general parameter to all translations is "target language," whereas, a specification for a particular translation instance could be "United States-English."

The translation specifications help guide the selection of error categories needed for the assessment. Once the needed error categories have been identified to assess the compliance of the translation to its pre-determined specifications, nearly all of the pieces are in place to begin the assessment. The core portion of the MQM error category hierarchy is shown in Figure 1, with the full hierarchy shown in Appendix A.[16]

[12] defined features of a particular translation derived from a generalized set of translation parameters
[13] general characteristics of a translation, such as target audience, register, delivery date, etc.
[14] ISO stands for International Standards Organization. See http://ttt.org/specs for more details on this particular standard
[15] American Society for Testing and Materials (ASTM) see: www.astm.org
[16] The MQM hierarchy will be updated in 2015, Figure 1 shows the MQM core from when I conducted this study in 2014.

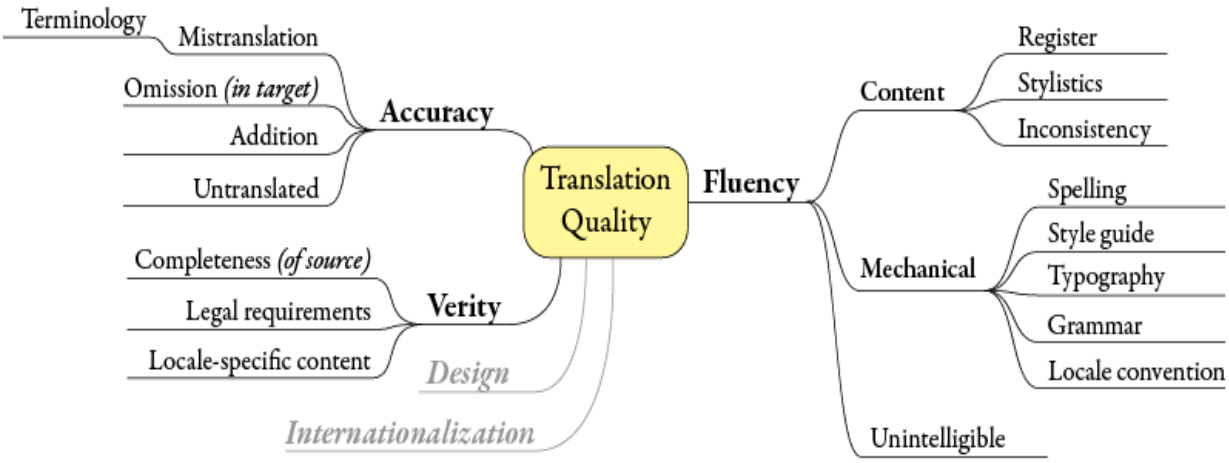


Figure 1: MQM core error categories (see Appendix A for the full hierarchy)

B. Aspects of translation quality assessment

Parameters, specifications, and error categories help to measure compliance to the unique requirements of a *translation* but how to best utilize those error categories in a translation quality *assessment* is determined by defining the goals and unique requirements of the assessment itself. This is done by describing the aspects of the assessment, a concept that Melby and I developed as part of this study. Translation assessment aspects are the consideration of, (B.2) Who does the translation quality assessment?, (B.3) What is assessed?, (B.4) Where is the translation quality assessed?, (B.5)When is a translation quality assessment conducted?, (B.6)Why is the translation quality assessment done?, (B.7)How is the translation quality assessment done?

Each one of these translation assessment aspects is described in detail here: www.ttt.org/tqbackground. For the sake of brevity I will only describe the translation assessment aspects as they pertain directly to this study.

A.1 "Who" does the Translation Quality Assessment?

There are several groups of people that can determine the quality of a translation. For this study I have chosen to focus on *Expert-based* assessment. Drugan, in her book "Quality in professional translation" identifies expert-based assessment as the most pervasive method currently used in the translation industry. Expert-based assessment, in a simple form, involves an experienced translator who is not the original translator of the text as the assessor. Sometimes a pair or even a group of experienced translators assesses the quality of the translation.

A.2 "What" is Assessed?

"What" refers to the focus of the quality evaluation or assessment. In many cases, the translation product or, in other words, the target text, is the focus of a translation quality assessment or evaluation. In other cases, the terms *translation quality assessment* or *translation quality evaluation* may refer to the assessment of something else entirely, such as the translation process, producer, or project. The focus of this study is on the translation *product*.

Sometimes in the translation industry, product refers to what is known as "service as a product." Translation companies provide various translation services to their clients, such as translation memory management, file format conversion, termbase management, and, of course, translation. Some companies view the various services they provide, or the total sum of all their services, as their translation product. The service-as-a-product way of thinking can be seen in the EN-15038 standard of translation quality[17], which does not include assessment of the target text, only the process used to obtain it. This definition of translation services as a product is confusing. For the purposes of this study, I define the translation product simply as the translated text. Any other service provided by a translation company I refer to as a "translation service."

Assessors and evaluators often want to determine whether a translation is ready to be sent on to clients, or perhaps they are more concerned with finding out whether a particular translator is skilled enough to take on future translations. In both scenarios, how well the final translation product adheres to the original specifications is often the focus of a translation product assessment or evaluation. This can be difficult, especially if the specifications are implicit or unstructured.

[17] http://en.wikipedia.org/wiki/EN_15038/0

A.3 "Where" is the Translation Quality Assessed?

Translation Sector

Some examples of a translation sector include, but are not limited to: academia, translator certification, industry, and government. Each of these areas has different reasons for doing translation work. Translators, assessors, and evaluators in each sector deal with distinct combinations of translation specifications, as well as quality assessment and evaluation aspects.

Industry is a sector of particular notice because of the high volume of translation output and monetary transactions when compared to other translation sectors. Although exact numbers are unknown, it is obvious that more paid translations are completed by those in industry translation than those in academia or translation certification, due to the large number of people involved in the translation industry. For these reasons, I have chosen to focus this study on translation quality assessment in the translation *industry*.

A.4 "When" is a Translation Quality Assessment Conducted?

"When," in this case, does not refer to the time of day the assessment is being conducted, but rather at what stage in the translation workflow the evaluation is being carried out. Translation quality assessment can be done at virtually any point in the

translation workflow. There can be an assessment of the initial translation; after a translation has gone through a revision process; or even an assessment of the source text alone before the initial translation, to check for writing inconsistencies and potential difficulties for a planned future translation. Considering when a translation is being assessed or evaluated influences factors such as the required skill set of an assessor or evaluator needed to complete an assessment or evaluation at a particular point in the translation workflow. For instance, translators sometimes are not experts on the subject matter they are translating. Sometimes draft or final translation products containing highly technical writing undergo additional evaluations by monolingual subject matter experts to assure that the translated information is factually correct. Since the subject matter expert is sometimes not fluent in the source language, the translation must be far enough along in the translation workflow to allow for linguistic comprehension by a monolingual subject matter expert of the material in question. For this particular study, I have chosen to focus on translations after *initial translation.*

A.5 "Why" is the Translation Quality Assessment Done?

For every translation situation, there are different reasons for wanting to complete and then assess the translation, and new ones are invented every day. There are also many combinations of ways to classify these reasons for conducting quality

assessment. Often the purpose of analyzing a translation is either to assess or evaluate its quality. Although the terms "assessment" and "evaluation" are used interchangeably in everyday speech, they are, in fact, distinct in meaning. Assessment is directly related to formative analysis tasks, whereas evaluation is related to a summative analysis. Because assessment is a formative process, the goal is to produce as much detail as possible during a translation quality analysis. In an evaluation, the goal is to determine if something has met the predetermined specifications to serve as feedback for improvement on future translations. As shown in the following example section, there are many use cases in translation that would call for either an assessment or an evaluation. Although the MQM framework was designed to assess or evaluate translation quality, in this study I analyze MQM specifically as an *assessment* tool.

A.6 "How" is Translation Quality Assessment Done?

The five translation assessment and evaluation aspects already explained help to determine the assessment or evaluation method, or "how" the assessment or evaluation is to be done. If one is already using a particular assessment method, the other five translation aspects help to determine whether another method might better suit a

project's particular requirements. Put simply, this section helps to define the general classification of the many existing translation assessment and evaluation methods.[18]

To explain this particular aspect of translation, we examine three sub-aspects related to the following questions:

1) Is the text being looked at as a whole, or segment by segment with additional focus on the sub-segment level as needed?
2) Does the method being used have multiple error categories or just one?
3) How are the error categories being used?

Holistic vs. Analytic:

Is the text being looked at as a whole, or segment by segment with additional focus on the sub-segment level as needed?

A holistic approach, like the name implies, considers a translation as a whole. Holistic approaches generally give a quick measurement of how well a document has been translated. This method provides a fast determination of quality. On the down side, if translators are looking for formative feedback on their work, a holistic determination of quality might not be detailed enough for focused translation improvement. In contrast, an analytic approach to measuring translation quality involves smaller portions of the text; e.g., at the word, phrase, or paragraph level,

[18] In this case "assessor" or "evaluator" could refer to any one of the "who"s previously mentioned.

allowing translators some indication as to where a translation error was identified by the assessor in the text.

Multidimensional or One-dimensional:

Does the method being used have multiple error categories or just one?

Texts or portions of texts can be examined for one particular type of error, or many different types of errors. These fall into error categories,[19] such as spelling errors, addition errors, omission errors, etc. Dimensions[20] constitute generalized classes or families of related error categories. Some of these include the general concepts of accuracy, fluency, and verity. Accuracy error categories have to do with errors that occur between the source and target texts, such as mistranslations or omissions. Fluency errors, on the other hand, can be identified in the source or target texts on their own, without the need to compare the two against each other. Some examples of fluency errors include a spelling error or grammar error in the target text, in which there is no need to look at the source text to identify that an error has occurred. Verity errors involve elements of the translation which pertain to the relevancy of the translation to the target audience. For example, in Spanish after a person sneezes, it is polite to say "salud." If they sneeze again within a short period of time, you say "dinero." If there a third sneeze, then the proper response is to say "amor." While this interaction could be

[19] Type or class of an error

[20] A dimension is a group of related error categories.

translated quite fluently and accurately into English as, first sneeze: health, second sneeze: money, third sneeze: love, the interaction would not be well understood by the target audience in that it is not a typical English cultural convention.

Whole dimensions may need to be considered or excluded in an assessment or evaluation, based on the needs of a particular translation. For instance, if a requestor desired to translate product surveys, and just needed to get a quick and cheap determination of whether or not the customer was satisfied or dissatisfied with the product, then they might rely on machine translation, a quick human translation, or a summary translation to get the desired outcome at the most affordable price. Fluency would not be a factor in the translation evaluation, whereas accuracy would be most important. Translation assessors and evaluators can focus on one general type of error (one-dimensional) or they can focus on finding several different classes of errors (multidimensional). Both methods are potentially useful for given situations. For example, an elementary school teacher grading spelling tests is only concerned with one particular error category, spelling. Other situations may call for a larger set of error categories.

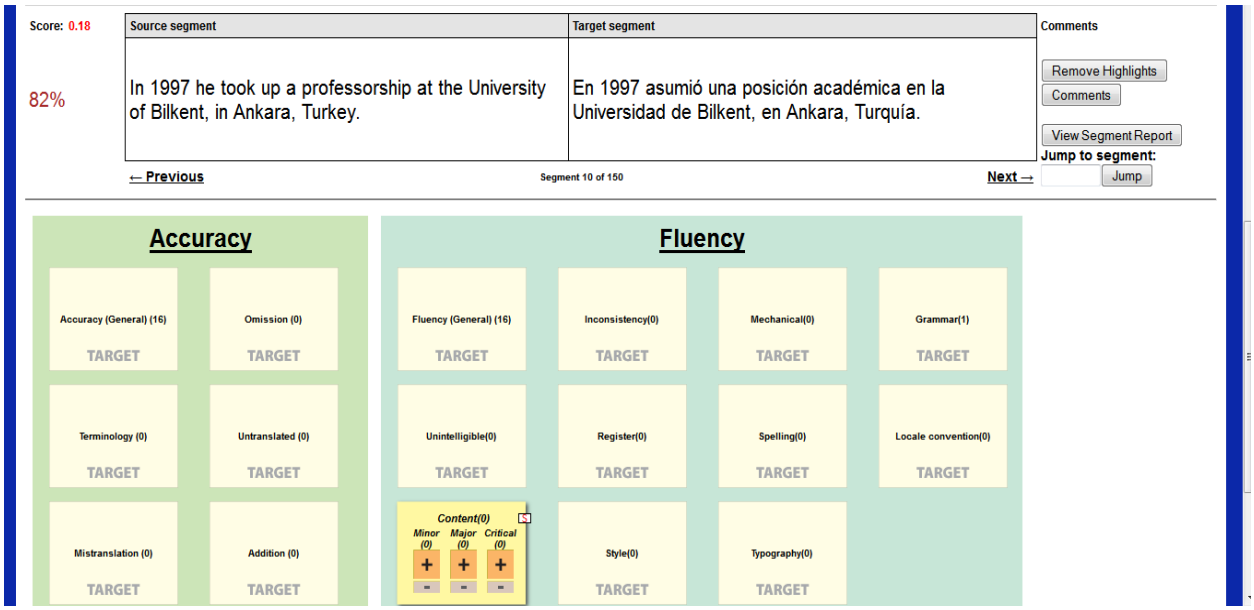


Figure 2: A scorecard on www.scorecard2.gevterm.net

Using translation specifications as a guide, any number of error categories can be organized into a metric,[21] which can then be rendered as a scorecard,[22] as shown in Figure 2, which is then used to measure the quality of a translation.

A scorecard is one implementation of the metric, and is similar to any typical scorecard used in other sectors—such as keeping track of strokes in golf—whereas the conceptual layout of that scorecard—such as how many holes and strokes make up par for a particular golf course—is most akin to the concept of a metric. In other words, the metric defines how a quality score is to be determined, and a scorecard records an

[21] A metric is a selection of error categories for the purpose of measuring translation quality against the translation specifications.

[22] A sheet, program, website, or item based on a translation metric that helps to record translation errors in order to determine translation quality. When referring to the integrated concept of both the scorecard along with its corresponding metric, the word metric is used.

actual quality score for a particular instance based on the metric. A translation quality assessment and evaluation framework can then be understood as a tool for the creation of translation quality assessment and evaluation metrics. Continuing with the sports analogy, the framework would have a list of categories such as strokes, baskets, goals, strikes, fouls, penalties, or any other concept needed to create a metric for determining the score of a sport. This would be rendered as a scoreboard which would display a few categories for determining the current status or outcome of a sport.

Error Category Use in Translation Assessment Metrics

How are the error categories being used?

There are several ways in which error categories are used in translation quality assessment and evaluation, four of which are explained here.

First, there is the pass/fail method, in which a whole translation or smaller portions of the translation are rated as either a pass or fail in spelling, grammar, accuracy, or any of the other of the numerous possible error categories. This method is sometimes used in traditional essay grading, where the whole text is rated pass or fail based on predetermined criteria.

Second, error categories can be rated on a scale. Assessors or evaluators are asked to rate the translation's spelling, for example, in the entire passage or a smaller portion of the text on a scale, e.g., one to five.

Third, another method is to rank translations in order of highest to lowest quality or vice versa, rather than by a specific error score. This is a type of normative assessment. It can be done with specific error categories as well, such as ranking translations according to fluency, for instance. Ranking only works if there are multiple translations available of the same text, which is most often not the case in industry translations.

Another approach, called the error marking approach, is when a translation quality assessor or evaluator simply marks all the errors he or she finds, counts them, and classifies the errors into their corresponding error categories. The overall percentage score, known as a quality score,[23] then helps to determine whether or not the translation is acceptable, by comparing it to an acceptance threshold.[24] For example, if a translation was evaluated in a highly-regulated industry, such as handling dangerous chemicals, the acceptance threshold would be very high, e.g., no errors, due to the intense need for correct handling procedures for both the source and target audiences. By contrast, translation done in an immediate emergency, such as trying to give someone proper directions to the hospital when someone is injured, might have a lower threshold as long as the translation is fast and accurate enough to accomplish the task.

[23] calculated by taking the total amount of errors times the severity of the error and dividing the total by the length of the text

[24] acceptable quality score before rejecting the translation

The acceptance threshold is determined by end-user needs and is derived from the translation specifications.

One more important factor in the creation of MQM metrics is a set of specific weights for individual issues and thresholds. These provide the interpretation of the metric for a specific case. For instance, in some cases there might be a well-defined set of terms that need to be used, wherein for that particular assessment terminology would have a much higher impact or higher weighting in the overall determination of quality.

Although MQM lends itself as a framework for creating metrics of any shape and size, I have chosen to analyze MQM in the form of an analytic-multidimensional-error-marking metric since this type of metric will give the most information possible during an assessment with the goal of recording as much formative feedback as possible to improve future translations.

B. Summary

To create an MQM-style metric, one is invited to define the specifications for the particular translation project, and then consider which translation aspects they are dealing with. The specifications guide the user in selecting error categories for the creation of a metric, and the aspects define how to use the metric, whether it is holistic or analytic, etc., which guides the user in implementing the metric as a scorecard. This

process of using aspects and specifications to create MQM metrics is illustrated in Figure 3.

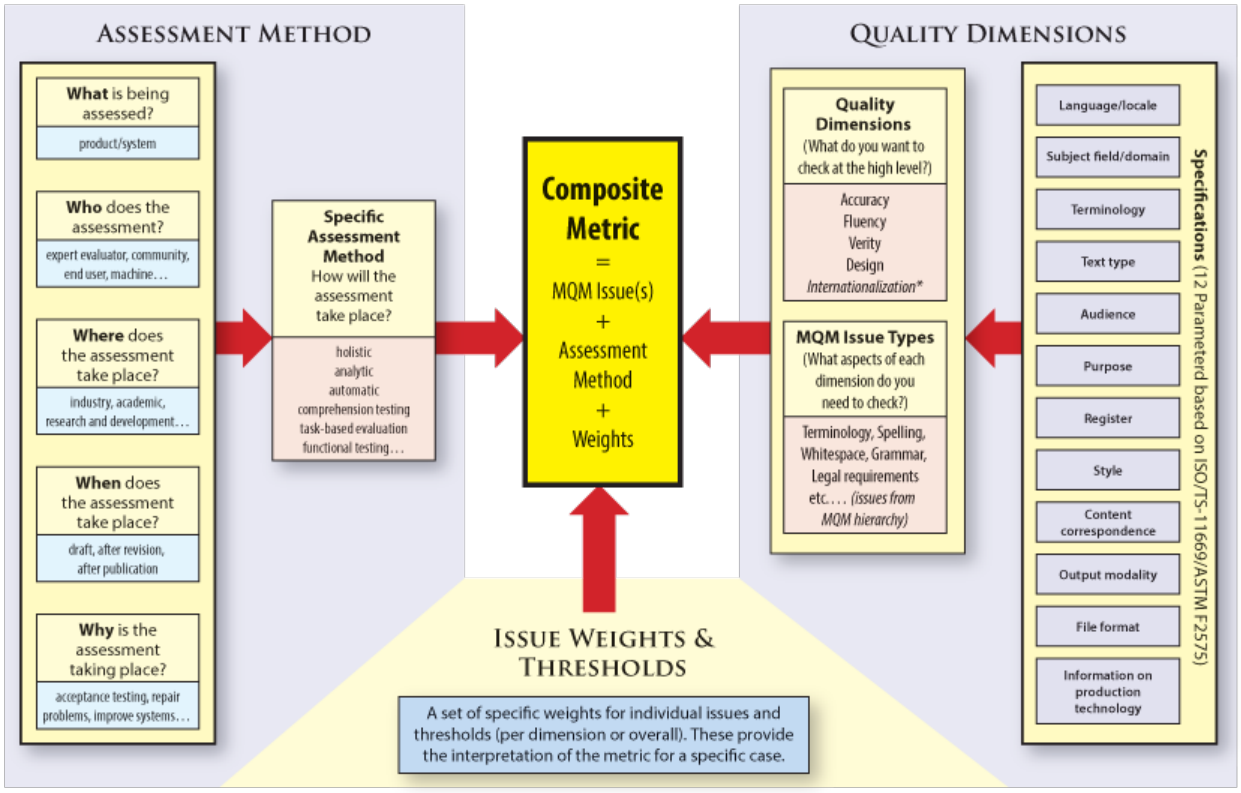


Figure 3: Creating MQM metrics (graphic courtesy of the QTLaunchPad Project)

The error categories and selected method combine to form a metric. The metric is then implemented as a scorecard. Once created, the scorecard can be used by assessors or evaluators to help determine the quality of a translation in regards to accuracy, fluency, verity, and stakeholder needs, as defined in the translation specifications. I will further explain how I will examine my chosen translation assessment aspects in the methodology section of this study.

VIABILITY

MQM is potentially a useful tool to assist in the assessment of the translation quality of industry translation products. In this study, I put the MQM framework to the test, to determine whether it is a viable solution based on the current needs of the translation industry. I do this by examining the reliability of assessors using specific MQM metrics, as well as the validity, and practicality of the MQM framework, and metrics derived therefrom.

A. Considerations for Determining the Viability of a Translation Quality Framework

A.1 Definition of Viability

Viability, in general vernacular, is understood as the functional usability and sustainability of something. When applied to a translation quality assessment and evaluation framework, viability is combination of assessor reliability using the metric, and additionally the validity, and practicality of the metrics produced from the framework.

In this study, I examine how reliable assessors are at identifying and classifying errors using the same MQM metric, while assessing the same translation. I assure the validity of the metric by establishing translation specifications with the help of the

translation requestor, which help to guide the selection of error categories to include in a metric.

A.2 Framework vs. Metric

Establishing the viability of a framework and of a particular metric made from that framework are two entirely different tasks. A framework is similar to a factory that manufactures a variety of tools, whereas a metric is like a tool made in that factory. An experienced craftsman could easily tell you what characteristics a tool would need for everyday use in his particular trade, as well as which tool he needs to use from one task to the next; but he would not know if every tool created by the factory could be used to fit every scenario he might encounter, without years of experience using the tools on the job.

Applying this example to translation quality assessment, it is relatively easy to determine the assessor reliability, validity, and practicality of a particular metric; but to do the same for an entire translation quality framework, such as MQM, is beyond the scope of this study. Much like a tool factory is only as good as the tools it can create, a translation quality framework is only as good as the metrics it can produce. To this end I begin to establish the viability of the MQM framework in this study by determining the viability of metrics created therefrom.

A.3 In Regards to Practicality

The practicality of a translation metric can be established in several ways; for instance, by considering the ease of use, monetary cost of implementation, or the speed of assessment or evaluation. Normally, practicality is very important when determining the viability of a translation quality framework. In the specific case of MQM, practicality cannot be easily determined due to the customization that the framework allows. For example, we might compare an MQM metric to a regular multidimensional, analytic, error-marking metric like SAE J2450. MQM can be configured to produce a metric that is faster to use, by making it holistic or by using fewer error categories than SAE J2450. Furthermore, MQM can also be configured to produce a metric that is less practical than SAE J2450, by doubling the number of error categories, for instance. Based on these considerations, when comparing MQM to other existing (non-customizable) metrics, MQM can be configured and optimized to produce a more practical metric in most cases. Even if one were to prove a single MQM metric to be impractical, one could redesign the metric into a more practical emulation.

Thus, by proving the practicality for a single MQM metric, one cannot directly infer the practicality of the MQM framework as a whole without many more similar studies. However, as in the toolkit example, we can rely on the extensive experience of those in the translation industry as a starting point to determine if the MQM framework is a complete framework for the assessment of translation quality in that sector. In the

next section, I describe a survey which was completed as part of this thesis study, in which 32 industry translation companies indicated the most common error categories required for determining the quality of the translations in their industry.

B. The GALA Survey

In 2012 and 2013, at least three independent surveys of industry translation practices were published. One was by Sharon O'Brien (2012) working with TAUS, in which they surveyed 11 of their member organizations to determine how quality assessment was being done in industry. "Quality in professional translation," by Drugan (2013), is a culmination of ten years of talking with translation companies on the topic of translation quality assessment and evaluation. In 2013, the Globalization And Localization Association (GALA), which is an organization consisting of hundreds of translation companies, content owners, and freelance translators, decided to survey their members on how they typically perform translation quality assessment and evaluation. Administrators from over 400 GALA member organizations from all around the world responded to the email survey. This is the survey that was mentioned in the background section of this study. The survey consisted of questions such as:

1) How often do you evaluate the quality of human translation?
2) Do you use any sort of translation specification / brief to instruct translators on your expectations and, if so, what sort?

3) Do you see the need to change your quality assessment processes in the next 2 years?[25]

Upon learning about the GALA survey, I was able to obtain the summary results through some contacts on the MQM team at DFKI (German Research Center for Artificial Intelligence), who were then able to put me in touch with Serge Gladkoff at GALA who had access to the survey results. The results of the survey are a comprehensive summary of current quality assessment and evaluation practices in the translation industry. I found that the survey did not ask some fundamental questions[26] needed for examining the viability of MQM against industry needs. Specifically, the survey did not ask anything in regards to error category use in the translation industry.

In the next section, I describe how I extended the GALA survey into a second phase, in which I was able to ask translation companies about their translation quality assessment and evaluation practices in regards to which error categories are commonly included in typical translation quality assessments in the translation industry.

[25] A full list of all GALA survey questions can be found in Appendix B

[26] The missing fundamental questions are described in the methodology section of this study.

METHODOLOGY

In this section, I discuss the survey and the case studies I use to establish the viability of the MQM. As part of this thesis study, I have extended the GALA survey into a follow-up survey, to discover which error categories are typically used in industry translation quality assessment and evaluation practices. The results of this survey are compared against the current set of MQM error categories. This demonstrates whether the MQM framework has a complete and valid set of error categories when compared to current industry needs.

The two case studies discussed in this section are the Firefox and MultiLing case studies. The purpose of the Firefox case study is to prove the validity and assessor reliability of one instance of a metric derived from the MQM framework. The MultiLing study is meant to further examine validity in a different form, by comparing the differences between an MQM-style metric and the already-established industry standard SAE J2450 metric. Due to financial constraints on the part of MultiLing, their case study was not carried to completion. Despite not obtaining final results, I continue to include the MultiLing case study, due to the many valuable insights discovered during its planning stages.

A. The GALA Follow-up Survey, Continued

After negotiation with initial survey stakeholders and following GALA privacy police guidelines, I obtained permission to conduct a follow-up survey. In that survey I asked the participants two additional questions beyond what was asked in the initial survey, along with a few other questions added by GALA. The missing fundamental questions which I chose to include in this survey focused on establishing the required error categories to determine the current quality assessment and evaluation needs in the translation industry. The questions are as follows:

- What error categories are the most important to check for? Perhaps name the top 3 most important.
- What are the most unusual categories you check for?

Before conducting the first follow-up calls, I obtained Institutional Review Board (IRB) approval from the Brigham Young University IRB board. The IRB is a university council that reviews students' and faculty's research projects and ascertains whether the research might be potentially harmful in some way to the subjects participating in the study. Approval was obtained quickly, due to the fact that the interviewees had voluntarily participated in the first survey, and had also indicated whether or not they would be willing to participate in further research. I agreed to limit my attention to those who indicated in the affirmative that they would answer further questions. This turned out to be 172 of the 400 participant organizations.

Those at GALA and I also decided that the follow-up survey would be conducted in a phone interview format, with a basic script that could be adapted during the calls as needed. We decided that the goal would be for me to interview at least 50 of the survey respondents over the course of several months. We later decided that 50 would be an ambitious goal to strive for, but that 30 would be the minimum acceptable number to obtain a 95% confidence interval within plus or minus 15% that the responses were an accurate representation of the entire participant population. I then sent an invitation to participate in the follow-up survey to all 172 previous participants. I quickly contacted any participants that responded with interest to the email invitation, and set up an interview with each of them. In the end, I was able to interview representatives from 32 total translation organizations. After two months of deliberation with GALA before the first interview call, I produced a set of interview questions that would be beneficial for both parties. The full script of the GALA follow-up survey, including my two additional survey questions, is shown here:

- What kind of translations does your company do?
- Your company mainly outsources your translations? What are some of the difficulties of maintaining quality that you find when outsourcing translations? (If applicable to company)
- What are some of the ways you are double-checking your translation quality?
- How much time is spent in the revision / quality assessment process compared to time spend on the translation itself?
- Your top language pairs are: [List language pairs]. Do those languages provide any particular problems with quality assessment?
- In the GALA survey, you indicated that your company uses known CAT tools for quality assessment. Could you please tell me about these? (If applicable to company)

- Are your quality assessment methods the same for all language pairs and all projects?
- What error categories are the most important to check for? Perhaps name the top 3 most important.
- What are the most unusual categories you check for?
- We imagine that you have some way to communicate your expectations to translators; how is this done? (Do you give any sort of translation brief to your translators?) May we have a copy?
- If your company uses machine translation, do you evaluate it differently than human translation? Could you please explain further about those differences?
- Looking to the future, would you be willing to be the contact person in your company for future GALA projects?

One supervisor from GALA, another from DFKI, and Melby oversaw my first call to one of the surveyed companies, and gave formative feedback on how I could improve future interviews by properly conducting the calls, and in what manner the questions ought to be asked. After the first call, I created a password protected website to record the results of each interview. All those supervising the survey, including those at GALA and my committee chair, Melby, had access to the website and the results of each call, which were available immediately after I uploaded them to the website.

I was not given permission by GALA or by the participants to record any of the conversations, but I was allowed to take notes during the interviews and write the responses to each question. During the calls, I asked each question in the interview script that was specifically relevant to the company, based on their answers to the previous survey, and I took notes as the participants answered the questions. I wrote a summarized answer to each question immediately after I heard it from the participants,

before moving on to the next question. After the call, I reviewed and revised my notes, adding additional information as needed, and then uploaded the results to the website. The people at GALA, DFKI, and Melby could then view the call results. If something was unclear to them, they would contact me for clarification. If I was unable to provide clarification, I was allowed to re-contact participants via email. The interview responses are still recorded on the database, but they are not yet publically available beyond the summary results presented in this study. The survey results work in tandem with a series of case studies in order to determine the viability of MQM.

B. Case Studies

For these case studies, I decided that the translation industry would be the focus of the subsequent case studies, because of the large volume of translations that are completed and assessed in this sector. I surmised that, by determining the extent of viability in the following scenario, the results of this study are of maximum benefit to those involved in the industry translation. The scenario was outlined in the background section of this study and is summarized here:

- Who: Expert-based assessment
- What: Translation products
- Where: Industry
- When: After final translation
- Why: Acceptance testing
- How: Expert assessment using a multidimensional analytic error-marking scorecard

B.1 Setup For Case Studies

In order to implement MQM for this combination of translation aspects, it was necessary to derive a usable metric and scorecard from the MQM framework. To this end, I developed a web-based scorecard tool, found at http://scorecard2.gevterm.net,[27] which allows anyone to create and use customized MQM metrics based on the ISO 11669 subset of translation parameters, to produce overall quality scores, and to classify particular errors found in a translation text. The tool was designed specifically for this thesis study and, therefore, cannot handle the full MQM framework. It is freely available to the public at the previously mentioned website. The only "how" sub-aspects currently available for use on the MQM scorecard web tool is the analytic error-marking method used for one-dimensional or multidimensional assessment designed for the aforementioned scenario.

I created the tool using a combination of the HTML, PHP, JavaScript, CSS, JQuery, and MySQL programming languages, all of which I learned while I was developing the web tool. To create a new metric, users must first answer a set of parameter questions to define their specifications. The website then directs project managers to a page that allows them to select MQM error categories. Once the

[27] Since development of my initial scorecard2, additional iterations have been developed by DFKI. Further improvements are planned in the near future. The current version can be found at http://scorecard4.gevterm.net.

specifications and error categories have been defined and selected, the web tool creates a metric that contains only the error categories that the user has selected. Users have direct access to the metric at this point, and are able to import the metric into their own account or share with other users as needed.

When the metric is rendered as an analytic-multidimensional-error-marking scorecard, all error categories are neatly displayed in their own sections on the web page. The webpage displays a description of the error category when users hover over the name of the error category. Each error category section contains buttons for adding or removing the number of errors found in the text that fit the error category description. A method for uploading a translation text is also included on the metric webpage. Once uploaded, users can view each source-target segment[28] of the translated text, one at a time, and are able to jump to any segment of the text at any time.

The metric page also records all activity, including when the page is opened, when an error is found or deleted, and allows users to add a comment to any translation segment for later reference. Input for the text length and other miscellaneous options are also available on the webpage. An overall quality score is updated and displayed every time an error is recorded on the metric. The score is calculated by adding up the number of total errors, times the error category weight, divided by the

[28] This is a portion of the source or target text; a word, sentence, phrase, or paragraph. Often segments appear in translation as bilingual translation units which are aligned so that the source segment and target segment are roughly equivalent in meaning.

length of the text. The result is a quality percentage score displayed near the top left corner of the scorecard. It updates every time an error is identified, classified, and marked using the scorecard[29].

B.2 Case Study: Mozilla Firefox

The first study is the assessment of the translation of the Mozilla Firefox web browser[30] and the assessment of the Mozilla Firefox mobile phone operating system. Both were previously translated from English into 11 different dialects of 7 different languages. The design of the assessment portion of the case study came from direct interaction with Mozilla. Having heard about my ongoing studies into the MQM standard and the development of the MQM metric-scorecard website from Melby, the head of localization at the Mozilla Corporation asked how they could get involved. He agreed to help by assisting me in setting up a case study, using the industry scenario found in the background section of this study, to assess the translation of their Firefox web-browser and mobile OS. I worked together with the Mozilla Corporation, using the MQM scorecard web tool to select the specifications and error categories that best suited the needs of this particular project, to assure the maximum validity of the metric used in the case study. The specifications for this case study are shown in appendix C. The error categories we selected for this customized metric are listed here:

[29] refer to Figure 2 for a screenshot of the scorecard on http://scorecard2.gevterm.net
[30] version 28

- Accuracy branch
  - Terminology
  - Mistranslation
  - Omission
  - Untranslated
  - Addition
- Fluency branch
  - Inconsistency
  - Register
  - Spelling
  - Typography
  - Grammar
  - Locale convention
- Verity branch
  - Verity general
  - Locale applicability[31]

The assessment was done using the same metric across all languages. Assessors were comprised of various bilinguals from around the globe, including professional translators who actively participate in Mozilla translation projects, as well as students of translation and linguistics from Brigham Young University. All participants were required to attend a one-hour training session, in which I instructed participants on the use of the customized Firefox metric and how to interact with that metric through the MQM web scorecard tool. The training was recorded for those participants who could not be there in person. The training included examples of translation errors and the corresponding error categories in which they should be classified. All participants had access to a searchable Mozilla translation memory database, called the Transvision

[31] For error category definitions, please refer to the MQM website at http://www.qt21.eu/mqm-definition.

database, found at http://transvision.mozfr.org, which contains all previous translations completed on any Mozilla project. Participants were instructed to use the Transvision database whenever they were unsure about a translation, instead of relying on other translation sources such as Google.

The Mozilla localization team and I divided the case study into four separate week-long sessions. Each week we completed a slightly different assessment, as outlined below. All assessors participated on a volunteer basis. Assessors during all sessions were allowed 30 minutes each day, Monday through Friday, to assess as much of the translation as possible. They were also allowed to either join the entire group of assessors in person at a computer lab on the Brigham Young University campus or join the group virtually through a web chat interface. The time that each person spent doing the assessment was recorded by the MQM web scorecard tool.

We did not give any incentive to the global participants from the Mozilla translation community of volunteers. Mozilla did instruct their community to add participation in the study to their normal work projects on a volunteer basis. Because of this, only five Mozillians participated in the case study. Both incentive and participation for Brigham Young University students was much greater. Mozilla offered student participants Firefox T-shirts, food for those who attended the sessions in person, and a grand prize of a new smart phone from Mozilla. Only those assessors who participated in the initial training and all five 30-minute sessions during the week were qualified for

the drawing to win the smart phone. Class credit was also received by some students as permitted by their professors.

The four sessions were outlined as follows:

- *Week 1* – Assessment of the Firefox browser version 28, English into Mexican-Spanish only (10 student assessors participated, though only 5 participated the full 5 days)
- *Week 2* – Assessment of the Firefox browser version 28, English into Parisian-French only (4 student assessors participated)
- *Week 3* – Assessment of the Firefox browser version 28, English into Mexican-Spanish, Portugal-Portuguese, Eastern-Armenian, Spain-Spanish, Chilean-Spanish, and Mainland-Chinese. (11 student assessors participated)
- *Week 4* – Assessment of the Firefox mobile operating system, English into international-Spanish, Mainland-Chinese, Parisian-French, International-Russian, International-Italian, and Taiwanese-Chinese. (20 student assessors participated)

Due to low participation for most languages, I decided to analyze only the language with the most participants from all four weeks. This turned out to be the translation of the Firefox browser into Mexican-Spanish. Assessor scores for the Mexican web browser translation were qualified against results provided by a panel of translation experts. These experts were from the Mozilla Corporation, supervisors of the teams that conducted the initial translations of the browser and operating system. Review of Firefox translations is a normal part of the supervisors' day–to-day work. The experts conducted an assessment of the translation with MQM, using the same scorecard tool as the other assessors. The results are presented in the data analysis section of this study, in which the results of the experts and the scores from the regular

assessors are compared to determine whether the assessors are reliable when using the MQM metric.

B.3 Case Study: MultiLing

After learning of my involvement in the development of MQM, the regional vice president of the Provo branch of the translation service provider MultiLing volunteered his company's services to test MQM. Unfortunately, time and financial constraints caused major complications for the study to proceed and we were unable to complete it. Still some results from the initiation of the project are relevant for this discussion.

For the case study, we decided to test MQM against the existing industry-established metric SAE J2450. MultiLing had already been using this metric to assess the quality of some of their translations, and they were interested to see how MQM scores might or might not correspond to their SAE J2450 scores. The team of quality management at MultiLing chose to use one source text translated into four target languages. The quality management team then asked the company's corresponding language teams for the four chosen languages to rate the various translations, first using the non-customizable SAE J2450 scorecard, and then again using an MQM scorecard. Since neither MultiLing nor I had access to a SAE J2450 scorecard tool we decided to configure my existing MQM scorecard web tool to emulate SAE J2450. In order to do this, I had to work closely with MultiLing to create a customized MQM scorecard that

used the same error categories as SAE J2450. Fortunately, SAE J2450 error categories were included in the development of the MQM hierarchy; only the names of the error categories had changed. There were some procedural changes needed for the use of SAE J2450 that I had to work into the scorecard web tool. One example is that SAE J2450 can only mark errors in the target text, whereas in MQM, errors can be marked in both the source and the target. Another is that each error category in SAE J2450 has a different weighting; e.g. a mistranslation error is worth two typography errors. Finally, as mentioned, the error category names differed slightly between the MQM framework and SAE J2450.

Error categories for the customized MultiLing MQM metric are listed here:

- Accuracy
  - Accuracy General
  - Terminology
  - Mistranslation
  - Omission
  - Addition
  - Untranslated
- Fluency
  - Fluency General
  - Spelling
  - Typography
  - Grammar
  - Register
- Verity
  - Legal requirements

The error categories for SAE J2450 using MQM error category labels are listed here:

- Accuracy
  - Accuracy General
  - Terminology
  - Mistranslation

    - Omission
    - Addition
- Fluency
    - Fluency General
    - Spelling
    - Typography
    - Grammar
- Miscellaneous[32]

The main difference between the two metrics is that the ambiguous miscellaneous category in J2450 is not used in MQM. Instead, it has been split in the MQM metric as Verity, Legal Requirements, register, and untranslated. If any additional error categories were needed during the assessment when using the MQM metric, then those missing categories would be added from the hierarchy instead of being lumped together in a miscellaneous error category.

B.4. Putting It All Together

To individually measure the practicality, assessor reliability, and validity of the MQM framework compared to the current industry need I use a simple rating scale ranging from one to five for each concept. I designed the scales so that a score of 3 in any section is roughly equivalent to current industry translation quality assessment methods; anything above a 3 is considered an improvement, and anything below that number would demonstrate that MQM is not a viable solution to current translation

[32] not actually an MQM error category

quality assessment needs in the translation industry. Although the scales in each section include descriptions for scores of 2 and 1, for the case of MQM, any score below a 3 in any section automatically renders the entire MQM framework not viable to meet the current industry needs for the assessment of translation products, even if high scores are demonstrated in other sections. This is because a score of 3 is roughly equivalent to performance shown by other assessment metrics used in the translation industry, and if it were to receive a score lower than a 3, it would indicate that MQM is not an improvement to what is currently available for assessment purposes in the translation industry. This being the case, if a section receives a score of 1 or 2, I would actually give the section a 0. In addition, if any section were to receive a 1 or 2, the overall viability score for the entire framework would be 0.

C. Reliability

Intra-rater reliability measures how often an assessor can produce a similar results under the same circumstances. By contrast, inter-rater reliability measures how well an assessor performs against other assessors who are analyzing the same item under similar circumstances. Intra-rater reliability is a good indication of how reliabile the assessor is, while inter-rater reliability helps make a determination as to both the metric being used and the assessors. Therefore I focus exclusively on inter-rater reliability in this study.

Inter-rater reliability is measured by comparing assessor results when using a metric and verifying those results against results from expert assessors. Using the results from the Firefox case study, I determined the reliability of seven assessors using an MQM metric by comparing the results obtained from regular assessors against expert scores obtained from a panel of native-speaking translators of the target language. Reliability of overall MQM quality scores are analyzed, in addition to the reliability of error identification and classification. If assessors of translation quality cannot accurately identify translation errors, then they cannot accurately classify those errors. This may lead to stakeholders in the translation industry spending time and money fixing translations that are actually correct, based on a faulty translation quality review.

The results of the reliability analysis are valuable for improving error category descriptions as training material in future versions of the MQM framework to promote enhanced reliability among translation quality assessors. The scale for determining the reliability of MQM is as follows:

> 5: The majority of multiple, trained assessors give scores within 5% of expert assessors' scores, and select the exact same error categories as expert assessors when rating the same translation.
>
> 4: The majority of multiple trained assessors give the same scores within 10% of expert assessors' scores, and select categories within the same MQM branch as expert assessors when rating the same translation.
>
> 3: The majority of multiple, trained assessors give the same scores within 20% of expert assessors' scores when rating the same translation.

2: The majority of multiple, trained assessors give scores within 30% of expert assessors' scores.

1: Assessors' scores appear random compared to expert assessor scores.

D. Validity

Correct selection of error categories guided by the use of translation specifications is the key to a valid assessment or evaluation of a translation product. As mentioned in the Mozilla Firefox case study, I met with the main translation stakeholder to determine the translation specifications for the project. The specifications were then used in collaboration by the stakeholder and I, to determine which MQM error categories needed to be included, and which needed to be omitted for the assessment of this project. This assured that the metric used was a valid measurement of translation quality for that instance. The scale for the validity section is as follows:

5: The metric contains conceptually all error categories needed to measure the quality of a translation based on the translation specifications.

4: The metric contains conceptually all of the needed error categories to measure the quality of a translation based on the translation specifications, but only because miscellaneous catch-all error categories are used to classify any unanticipated error types.

3: The metric contains conceptually most of the needed error categories to measure the quality of a translation based on the translation specifications, but is missing one or two needed error categories.

2: The metric contains conceptually some of the needed error categories to measure the quality of a translation based on the translation specifications, but is missing more than three needed error categories.

1: The metric is missing a large number of the needed error categories to measure the quality of a translation based on the translation specifications.

In addition to the validity of a single MQM metric, validity is the only section in which one can make a direct inference as to the completeness of the entire MQM framework. In this study, I determine the validity of the MQM framework at large against industry needs, as indicated by the results in the GALA survey follow-up calls. If the framework has sufficient error categories to cover the varying translation needs in industry, then conceptually, MQM is able to produce equivalent metrics to cover those needs.[33] Both the most common and most unusual error categories, as indicated by the participants of the follow-up calls, must be represented in MQM by equivalent error categories. The scale for establishing the validity of the MQM framework against industry needs is as follows:

5: The MQM framework contains all error categories needed by the translation industry.

4: The MQM framework contains at least 75% of all error categories needed in the translation industry.

3: The MQM framework contains at least 50% of all common and uncommon error categories needed in the translation industry.

2: The MQM framework contains at least 25% of the error categories needed in the translation industry.

[33] But, a metric might not be valid if it does not contain sufficient error categories, even if the MQM framework does have all the required error categories.

1: The MQM framework is missing a majority of the error categories needed in the translation industry.

E. Practicality

Normally, for a practicality test of a specific metric, one would compare the amount of time it takes to use an MQM metric against the time it takes to assess the same text, using a more well-known metric such as SAE J2450. As explained in the MultiLing case study outline, I was able to create an MQM-style scorecard that functioned in essentially the same manner as SAE J2450. Based on this level of customizability of MQM, we see that MQM can be configured to produce a metric that could be either equal to, or faster to use than any metric that is currently available to the translation industry by limiting the number and type of error categories in the metric.

I also acknowledge that an MQM metric could be created that is slower than conventional metrics. If one were to conduct a study on practicality for an MQM metric and happened to find that the metric they were using was much slower than conventional metrics, then it would not rule out MQM as a viable framework. This is because a slow metric based on the MQM framework could be reconfigured for optimized performance by eliminating unused and distracting error categories. Therefore a practicality study on an individual MQM metric would not speak as much as to the framework itself. Therefore, when correctly applied, MQM inherently

produces more practical metrics than non-customizable translation quality frameworks currently available in the translation industry.

For the benefit of those wishing to conduct future practicality studies on individual translation quality assessment or evaluation metrics, I include a scale here which I do not use in this particular thesis study:

> 5: Metrics based on the translation quality framework are significantly faster than other similar metrics based on another translation quality framework.
>
> 4: Assessors using metrics based on the translation quality framework are slightly faster when compared to assessors using metrics from other translation quality frameworks.
>
> 3: There is no difference in terms of speed of assessment when using a metric based on the translation quality framework in question when compared to the speed of assessment using metrics from other translation quality frameworks.
>
> 2: Assessors using a metric based on the translation quality framework in question are slower than assessors using other translation quality frameworks.
>
> 1: Metrics based on the translation quality assessment framework are so impractical that an assessment cannot even be completed.

F. Establishing Viability

Using the scales from each section and taking into consideration the follow-up to the GALA survey, in tandem with the case studies presented, I determine whether the MQM framework is a viable solution for the assessment of translation products found

in the translation industry. If MQM receives a score of 2 or 1 in either the reliability or validity sections, then MQM in its entirety cannot be considered a viable solution for current industry needs, even if a high score was received in another section. A score of 3 in any section indicates that MQM is not necessarily an improvement on current industry practices but is as viable as other frameworks currently being used. Finally, a score of 4 or 5 in any section indicates an improvement in that section when compared to current industry quality frameworks.

DATA ANALYSIS

A. GALA Follow-up Survey Results

After talking to representatives from 32 different translation companies, I learned a number of valuable insights regarding the current status of quality assessment and evaluation procedures in the translation industry. Most notable is that no two translation companies have exactly the same approach to determining translation quality. One company did zero quality assessment; assuming, without verification, that their translators were of the highest caliber. To that company, any sort of assessment or evaluation was seen as a sign of mistrust towards the translators who worked for them. In contrast, I had the opportunity to interview a company who had developed an in-house, LISA-based, customizable, multidimensional analytic error-marking metric based on specific company needs used to evaluate all translation products. Other companies did not bother to do an assessment and instead performed revisions in which they just fixed the translation, making whatever changes they deemed necessary.

Only five of the 32 companies (about 15%) indicated that they were satisfied with their current quality assessment and evaluation practices. Twenty-three (about 70%) specified that they "see room for incremental improvement (e.g., better software, standard metrics)." Finally, three of the 32 companies (about 9%) said that they saw an urgent need to improve their current quality assessment and evaluation practices. These

results corroborate and confirm Drugan's findings, that companies do not have a consistent method for determining translation quality.

When asked, "How much time is spent in the revision/quality assessment process, compared to time spent on the translation itself?" companies tended to fall into one of two categories. One group of companies claimed that 10% of their time working on a translation project was spent doing quality evaluation. These companies, in most cases, were focused on finalizing their translation product to go out to the end-user. Companies in this category favored more holistic, summative evaluation methods. The second group of companies claimed that 30% of project time was put towards translation assessment, favoring analytic formative methods. This second group was concerned not only with fixing errors, but also finding out more about them by using analytic and multidimensional assessment methods. It was not clear from the survey whether companies in this category spent extra time on assessment for the purpose of giving formative feedback to translators, or for other reasons. Other less-common responses included, "I do not spend any time on quality assessment," and "I have no idea."

While many companies, especially those mentioned in the first group, are only concerned with fixing translation problems to get the translation to the end-user as soon as possible, other companies desire to correct the source of the encountered translation problems, with a more in-depth analysis of the translation using translation quality

assessment metrics. In the survey, 95% of all companies who claimed to use a metric identified that they used a custom, in-house metric based on the standards: EN15308, ISO 9000, and the LISA QA Model (GALA, 2013). Although often cited as the basis for the creation of customized in-house metrics, EN15308 and ISO 9000 are not translation quality frameworks for the creation of translation quality metrics. Instead they offer standards and general recommendations for operational procedures and policies. EN15308 does name a few example error categories that could be used in the creation of a metric, but it does not offer a standardized methodology for doing so. The LISA QA Model, in contrast, is a standardized metric with error categories and the ability to produce a quality score. All companies using the LISA QA Model unanimously admitted that they were not using the model as designed, but instead were using a modified version that they developed in-house. For companies using a metric they did not develop it in-house, SAE J2450 was often mentioned as their metric of choice. Both SAE J2450 and the LISA QA Model were incorporated into the development of the MQM framework. Because MQM was created in part by incorporating many diverse metrics, including the SAE J2450 and the LISA QA model, I believe that current users of both SAE J2450 and LISA QA Model metrics will find the use of MQM-style metrics to be similar to what they have already been working with.

To determine the viability of the MQM framework, a list of both the most common and uncommon error categories was compiled for use in the validity portion of the

viability scales. The error categories in the following list are in response to the survey interview questions as discussed in methodology section A of this paper.

Both lists below represent the complete set of answers given by all 32 translation companies, ranked in order by decreasing frequency.

Most common error categories:

- Terminology
- Accuracy
- Fluency
- Omission
- Target grammar
- Mistranslation
- Typographical
- Spelling
- Completeness
- Legal requirements

Most commonly mentioned unusual error categories:

- Offensiveness (Not in MQM)
- Tone
- Register (Formality)
- Style guide compliance
- Length of text
- "Do not translate"

As shown, "Offensiveness" is mentioned by the 32 interviewees as an error category. Notably, it was not an error category in the MQM framework at the time of the study but has since been added as a result of this survey. This is because before, MQM handled offensiveness not with an error category, but with a concept called severity. When using MQM, any error identified by an assessor in a text is classified

into an error category, and then additionally classified as a minor, major, or critical error. Officially in MQM, minor is defined as "the issue does not impact usability of the text;" major is defined as "the issue leaves the text usable but is an obstacle to understanding;" critical is defined as "the issue renders the text unusable." Other frameworks, such as SAE J2450, only have two severity categories, minor and major. In this study, I do not examine if assessors can reliably agree on error severity classification. Participants of the follow-up survey that mentioned "offensiveness" as an error category do not think that severity is a sufficient concept to capture errors of this type. As a result of this study, MQM developers at DFKI have updated the hierarchy to include offensiveness as an error category. This change will be available in the 2015 MQM hierarchy update.

Using the full MQM hierarchy of error categories shown in Appendix A of this study and the interview question results, I determined that MQM covers approximately 93% of the error categories listed here from the 32 translation companies. This puts MQM at a 4 out of 5 in the framework validity portion of the viability scales.

B. Mozilla Firefox Case Study Results

B.1 Reliability of the Mozilla Firefox Metric assessors

There are two considerations in determining the reliability of assessors using the MQM metric. First is the reliability of the *overall* MQM scores, and the second is the reliability of the *error category selection* between the tested assessors and a group of expert Firefox translation assessors. Table 1 shows the overall scores for the Mexican-Spanish Firefox desktop translation given by all assessors and additionally by the expert assessors.

Assessor	Overall Score Given	Difference From Experts	Reliability Category
Assessor 7	99.1%	21.1%	2
Assessor 2	93.1%	15.1%	3
Assessor 3	89.6%	11.6%	3
Assessor 1	88.2%	10.2%	3
Assessor 6	82.4%	4.4%	5
Assessor 5	73.8%	-4.2%	3
Assessor 4	63.2%	-14.8%	3
Average Overall Score	84.2%		
Expert Score	78.0%		
Difference Average Overall Reliability Rating	6.2% 3		

Table 1: Overall Mozilla evaluation scores

The overall scores demonstrate that most assessors tended to be less harsh when compared to the expert team. There is a wide gap between the extremes of

assessor scores. Assessor 7 gave the text a near perfect rating, whereas Assessor 4 gave the text a score 36% lower. The difference in overall scores between the average assessor score and the expert assessment team was 6% which is less than the 10% required for a 4 on the reliability scale.

> 4: The *majority* of multiple trained assessors give the same scores within 10% of expert assessors' scores and select error categories within the same MQM branch as expert assessors when rating the same translation.

As per the above definition, the determining factor for establishing the extent of reliability for the assessors using MQM is not the average, but instead the majority of scores given by multiple assessors. The last column in Table 1 demonstrates that the majority of assessors did not score within 10% but did score within 20%, of the experts. This means that even before examining the assessors' error category selection reliability, the overall MQM scores limit the highest possible reliability score to a 3, based on the viability scales criteria as described below:

> 3: The majority of multiple, trained assessors give the same scores within 20% of expert assessors' scores when rating the same translation.

Since error category selection is not a factor at a level 3 for reliability, it is not applicable for this particular analysis in determining the viability of the MQM metric. However, I present the error category selection reliability data and associated discussion here for the benefit of future research, as it adds additional insight into the reliability of overall quality scores when using MQM.

To fully understand the error category selection data from the Mozilla Firefox case study, we must first discuss two important questions relating to translation quality assessment.

1) Were the assessors able to correctly identify translation errors?
2) Were the assessors able to correctly classify those translation errors using MQM?

The ability to identify a translation error in a translated text is prerequisite to the ability to classify that error using a translation quality metric. Inexperienced assessors or evaluators tend to find errors in a translation that, upon closer inspection, are not errors at all. This is called a type I decision error.[34] In addition, untrained assessors and evaluators can also fail to identify existing errors in a translation; this is known as a type II decision error. Neither type I nor type II decision errors are useful for determining the translation error classification reliability of assessors when using a translation quality assessment metric. This is because to *classify* a translation error, one must first *identify* a translation error. For example, an assessor finds an error in a translation that turns out to not be an error at all; it would not matter if the assessor was using SAE J2450 or an MQM metric because, in either case, the error would be falsely classified, since the supposed error does not really exist. If an assessor failed to identify an error, then the assessor would not attempt to classify it either. In decision theory, type II errors are much worse than type I. Take the example of a physician diagnosing that a person is

[34] Not to be confused with a translation error, type I and type II errors are used here in the sense used in statistical hypothesis testing and decision theory.

sick with a disease, but later finding out the test gave a false positive (type I), versus a physician not diagnosing a disease which was actually present (type II). The severity of type II errors also applies the field of translation, although perhaps not as severe in most cases. Due to type I or type II errors, translation producers might spend a lot of time fixing errors that do not truly need to be fixed, or they might fail to fix errors that are present in the translation.

Tables 2 and 3 show the percentages of the number of correctly identified errors in the translation, type I errors, type II errors, and errors that were correctly identified but not accurately classified.

Assessor	Correct	Type I decision errors	Type II decision errors
Assessor 1	2	4	39
Assessor 2	3	6	36
Assessor 3	1	5	38
Assessor 4	5	29	28
Assessor 5	8	42	24
Assessor 6	10	13	31
Assessor 7	0	3	42
Average:	2	14.57	34
Average Compared to experts	22%	33%	79%

Table 2: Number of translation errors identified, type I, and type II decision errors

Assessor	Difference in error is close (same MQM branch)	Error identified but incorrectly classified
Assessor 1	1	1
Assessor 2	1	3
Assessor 3	1	3
Assessor 4	2	8
Assessor 5	2	9
Assessor 6	2	1
Assessor 7	0	1
Average:	1.28	3.71
Compared to experts	3.0%	8.6%

Table 3: Number of misclassified errors

The data shows that, on average, 33% of all errors found by assessors were type I, when compared to the errors found by the experts. In addition, compared to the errors found by expert assessors, Assessors 1–7, on average, failed to identify 79% of the existing errors in the text. Although most of the assessors failed to identify a high number of the errors in the text, the large amount of type I decision errors affected the overall quality scores to appear much closer than they really were to the experts' score. This is why overall quality scores can be deceptive. Specific instances of type I and type II decision errors are presented here.

Type I:

Segment 17

Source	Target
update channel.	*actualizar canal.*
	actualiz+ar canal
	"update+inf channel"
	"to update channel"
Assessors: Assessor 1: Critical mistranslation Assessor 2: No error found Assessor 3: Critical mistranslation Assessor 4: Minor grammar Assessor 5: No error found Assessor 6: No error found Assessor 7: No error found	Firefox experts: No error found

Table 4: Example 1 of a type I decision error

Assessors 1 and 3 most likely were not familiar with the use of *actualizar* for the English term "update." Assessor 4 most likely was expecting a conjugated form of *actualizar,* perhaps a command or subjunctive form as it is in English. Assessor 4 might not have been aware of the polite command verb form in Spanish where the verb is left in the infinitive. This form is especially common when a Spanish command is not directed at any one particular person. This type of error could be avoided by both translators and assessors if the term were properly listed and available in a termbase, which is basically a list of terms and how they should be translated.

Segment 152

Source	Target
Pull down to show history	*Arrastrar para mostrar el historial*
	Arrastr+ar para mostr+ar el historial
	"pull towards the ground+inf in- order- to show+inf the(gender.m) history-log"
	"to pull towards the ground in order to show the history log"
Assessors: Assessor 1: No error found Assessor 2: No error found Assessor 3: No error found Assessor 4: No error found Assessor 5: Minor mistranslation Assessor 6: Minor mistranslation Assessor 7: Minor mistranslation	Firefox experts: No error found

Table 5: Example 2 of a type I decision error

Arrastrar also has the meaning of "moving something across a surface," which means it is very common to hear it used for the English verb "to scratch." Since Assessors 5–7 marked minor errors only, it is more likely the case that the false error dealt with the word *historial,* which is an easy mistake for non-native Spanish speakers. The difference between *historial* and *historia* in Spanish is "a log of occurrences" versus "story" in the general sense, written or unwritten. History in a web browser is specifically a log of past sites visited on the web. Therefore, *historial* is probably the more correct term in Spanish for this particular scenario, although *historia* might still be an acceptable alternative.

Segment 376

Source	Target
Move to Group	*Mover a Grupo*
	Mov+inf a Grupo
	"move+inf to Group "
	"to move to group"
Assessors: Assessor 1: No error found Assessor 2: No error found Assessor 3: No error found Assessor 4: Minor typography Assessor 5: Major typography Assessor 6: No error found Assessor 7: No error found	Firefox experts: No error found

Table 6: Example 3 of a type I decision error

Typography refers to punctuation errors. Lack of punctuation is common in web browser user interfaces. In this case, the Spanish follows the same capitalization and lack of punctuation as the English. Therefore, there was no error present in the Spanish translation for this particular example.

Type II:

Segment 58

Source	Target
To stop private browsing, you can close this window	*Para dejar la navegación privada, puedes cerrar esta ventana*
	Para dej+inf la navegación privad+ gender.f ,pued+2 nd .per. singular. Cerr+inf est+gender.f ventana
	"in-order-to leave+inf the (gender.f) navigation (browsing) private+gender.f , can+2nd .per.singular close+inf this+gender.f window"
	"In order to leave the private navigation, you can close this window"
Assessors: Assessor 1: No error found Assessor 2: No error found Assessor 3: No error found Assessor 4: No error found Assessor 5: Minor Grammar Assessor 6: No error found Assessor 7: No error found	Firefox experts: Minor mistranslation

Table 7: Example 1 of a type II decision error

This was a case where the experts, being native Mexican-Spanish speakers, preferred the more direct translation of the English "to stop" as *terminar* "to terminate" or *parar* "to stop," rather than *dejar* "to leave." Although *dejar* could still technically be used in this context, it was not a direct translation of the English text. I am not sure what grammar error Assessor 5 found in this segment, as no other assessor, including the expert team, found anything wrong with the grammar in this translation segment.

Segment 127

Source	Target
Zoom	*Tamaño*
	Tamaño
	"size"
	"size"
Assessors: Assessor 1: No error found Assessor 2: No error found Assessor 3: No error found Assessor 4: No error found Assessor 5: Minor inconsistency Assessor 6: No error found Assessor 7: No error found	Firefox experts: Major Mistranslation

Table 8: Example 2 of a type II decision error

"Zoom" in English on its own is more akin to "zoom in." There is a sense of making something appear bigger or "getting closer." Also in this case, zoom in context is more of a verb, commanding the web browser to perform an action. Better translations would have therefore been *enfocar* "to focus," *engrandar* "to make bigger," *aumentar* "to augment," or *acercar* "to get closer." Only Assessor 5 correctly identified this as an error, but the classification given was incorrect. An inconsistency error is better defined as two correct translations being used interchangeably, such as *aumentar* or *enfocar* being used for "zoom." In this case *Tamaño* is simply a mistranslation of the English source.

Segment 302

Source	Target
Reset	*Inicial*
	Inicial
	"the start" or "the starting place"
	"the start"
Assessors: Assessor 1: Minor mistranslation Assessor 2: Major mistranslation Assessor 3: No error found Assessor 4: No error found Assessor 5: Minor mistranslation Assessor 6: Minor terminology Assessor 7: No error found	Firefox experts: Major Mistranslation

Table 9: Example 3 of a type II decision error

In this example, only Assessors 3, 4, and 7 have made type II errors. "Reset" is most often an English noun, meaning to set again or to make something how it was in the beginning. The noun form refers back to the action of setting again; it does not mean any specific time or place, as "the start" or "the starting place." In this case, where we are commanding the computer to do something, the correct translation is probably something more like *reiniciar* "to reinitialize." Half of the assessors in this case also recognized this discrepancy. Assessor 6 is close to the right classification, since terminology is a subset of mistranslation in MQM, but incorrect in this case, since a terminology error occurs when a translation is correct, but does not conform to a predetermined set of approved translations for the particular context.

Possible reasons for such high numbers of type I and type II errors

The enormous discrepancy between regular and expert assessors is possibly due to a number of factors:

1) The regular assessors were not native speakers of the target text, whereas the expert assessors consisted of a team of one native English speaker and two native Mexican-Spanish speakers.
2) Regular assessors had only one hour of training on the use of the MQM metric and its error categories. In contrast, the experts had access to the same training and at least one team member with extensive MQM experience
3) The regular assessors consisted mainly of college-level translation students, whereas the expert team was composed of professional translators who interact with the Firefox Desktop translation from United States-English into Mexican-Spanish on a regular basis. In other words, the regular assessors were not "expert assessors."

Referring back to Tables 2 and 3, since about 79% of the data consists of type II errors, we can only deduce the reliability of error classification from the remaining 20%. Compared to that percentage, only 9% of the total errors found by the regular assessors were actually identified and correctly classified, when compared to the errors found and classified by expert assessors. Three percent of errors were classified by regular assessors into similar branches of the MQM hierarchy, and 8% were correctly identified but erroneously classified into another branch of the MQM hierarchy. This information demonstrates that in 60% of the cases when an error was correctly identified, it was also correctly classified into the appropriate error category, or at least an error category in a similar MQM branch. This would give MQM assessors low marks on the reliability scale, even when we restrict the analysis to only the 20% of relevant data. Here is an

example of agreement between regular and expert assessors, taken from the 20% of relevant data:

Segment 151

Source	Target
Right-click or pull down to show history	*Clic secundario o arrastrar para mostrar el historial*
	Clic secundario o arrastr+inf para mostr+inf el (gender.m) historial
	"Click secondary or pull towards the ground+inf in-order-to show+inf the (gender.m) history-log
	"Click secondary or to pull towards the ground in order to show the history log"
Assessors: Assessor 1: No error found Assessor 2: Minor mistranslation Assessor 3: No error found Assessor 4: No error found Assessor 5: Minor mistranslation Assessor 6: Minor mistranslation Assessor 7: No error found	Firefox experts: Minor mistranslation

Table 10: Example of error classification agreement

In this case the three assessors that actually found the error also correctly classified it, when compared to the expert Firefox assessors. *Clic secundario* "secondary click" was unfamiliar wordage for the expert team. They preferred *Clic derecho* "right click."

B.2 Validity

As explained in the methodology section, I assured maximum validity by consulting directly with the stakeholder at the Mozilla Corporation to determine the translation specifications and needed error categories for the assessment of this particular translation. Setting aside the fact that, in the majority of instances, errors were not correctly identified by the assessors, there was not a single reported error which the assessors were unable to classify into an available error category. This indicates that the assessors viewed the metric as containing a complete set of error categories for this particular assessment project. The result puts the validity of the MQM metric at a 5 out of 5 on the validity scale, in the eyes of the assessors.

C. MultiLing Case Study Results

Though the Multiling case study remains uncompleted, there is one result to report from the initiation phase, in which I was able to emulate SAE J2450 using MQM error categories. This demonstrates the adaptability of the MQM framework to create metrics to fit varying needs. The study also shows which MQM error categories are needed if future researchers would like to compare SAE J2450 to MQM.

CONCLUSION

Through the means of this study I achieved several outcomes. I interviewed 32 different translation companies about their translation quality assessment practices and found that the MQM error category hierarchy includes all but one of the common error categories needed in the translation industry. This category has since been added to MQM. The survey results indicate that MQM contains a conceptually valid set of error categories needed for quality assessment in the translation industry.

In order to test the validity and inter-rater reliability while using MQM metrics, I created the first publicly available, online scorecard application which is capable of creating customizable-analytic-multidimensional-error-marking scorecards using the MQM definition.[35] I used the scorecard application to emulate the SAE J2450 metric for the MultiLing case study and to create a scorecard designed to assess the translation quality of the Mozilla Firefox web browser against its specifications which I implemented by using the scorecard application. I demonstrated that the scorecard application, and by extension MQM, can be used to create valid translation quality assessment metrics and scorecards.

From the results of the Mozilla Firefox study, I discovered that using MQM metrics requires considerable expertise in order to maximize inter-rater reliability. Interestingly the results demonstrate low levels of inter-rater reliability for error

[35] www.scorecard2.gevterm.net

identification and much higher reliability for error classification using MQM. This indicates that the task of error identification is much less intuitive than error classification, at least for the assessors in this experiment. I discovered that efforts to improve both assessment tasks, through the development of additional training materials and future case studies are a critical issue. Additionally, it is my opinion that properly trained native speakers of the target language will most likely outperform non-native speakers with the same amount of training when it comes to translation quality assessment. All assessors need to be pre-qualified in some way before undergoing training. One way to assure assessor qualification would be to present assessors with a brief training on error identification and then with an exam. Those assessors that pass would then be deemed qualified for more extensive training on error classification using MQM.

In this study, I presented the concept of "viability" of translation quality assessment metrics, and showed that viability could be demonstrated through the inter-rater reliability, validity, and practicality of a metric. I determined that the Mozilla Firefox MQM metric is viable but could be used more reliably than was demonstrated in the case study by involving better qualified assessors.

In addition, collaborating with Melby during this study, I developed the concept of translation aspects which include the who, what, where, when, and why questions that determine how an assessment takes place. This was a direct contribution to the

development of the MQM framework and provides a common vocabulary for the description and classification of translation quality assessment metrics.

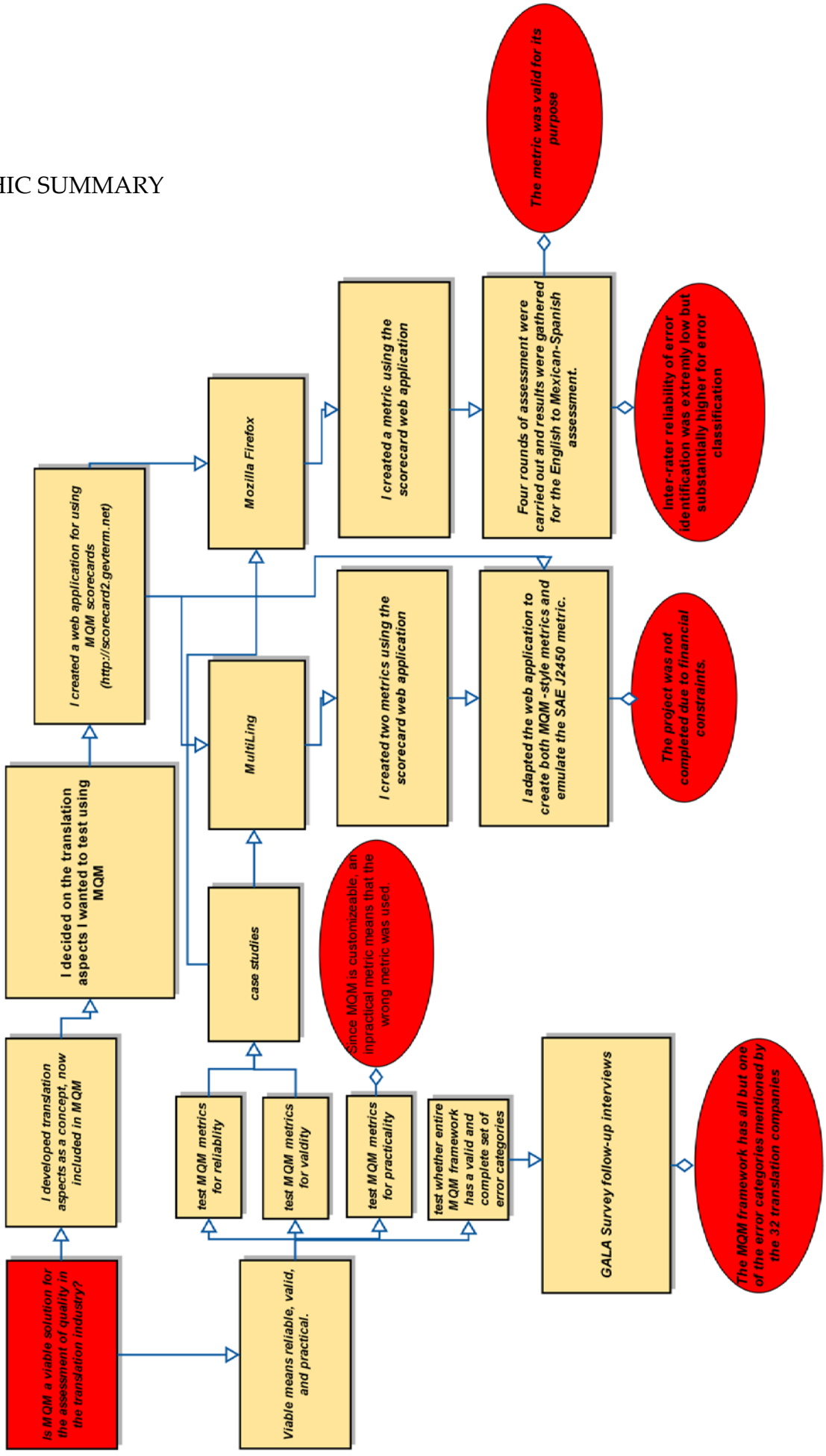
Based directly on the work I completed during this study, several questions on translation quality have emerged as a result. These include but are not limited to:

- How can we increase rater reliability while using MQM metrics?
- How can we pre-qualify assessors?
- How does the viability of MQM compare against the viability of other frameworks out there, such as the TAUS DQF?
- To what extent is MQM viable in different sectors outside of the translation industry such as in academia or government?
- Which translation parameters most directly affect the various MQM error categories?

Each of these questions could be the basis for future case studies.

In summary, the results I obtained from this study demonstrate that MQM is indeed a viable solution for industry needs, but the results also highlight the need for additional studies into the proper use of MQM, leading to highly-refined user guides and training materials to promote increased reliability among assessors. Overall, MQM is a viable approach to translation quality assessment that is ready to be used and further refined by those in the translation industry.

GRAPHIC SUMMARY



GLOSSARY OF TERMS

- **Accuracy**: the degree of correspondence of meaning interpreted from the source and target texts (i.e., does the reader of the target text understand the same content as a reader of the source text would?)
- **Analytic**: considering the translation in smaller pieces, such as word by word, sentence by sentence, or paragraph by paragraph
- **Assessment**: review of a translation for the purpose of formative improvement of translation processes
- **Assessment workflow**: Also called an assessment process, a sequence of translation quality assessments
- **Broad definition of translation**: includes localization and transcreation; Translation is the creation of target content that corresponds to source content according to agreed-upon specifications.(Melby et al, 2014a)
- **Content:** text plus optional non-textual elements such as graphics, sound, and color etc.
- **DFKI**: German Research Center for Artificial Intelligence (http://www.dfki.de)
- **Dimension**: A dimension is a group of related error categories.
- **End-User**: the target audience for a translation
- **Error**: a specific instance of an issue that has been verified to be incorrect
- **Error category**: a type of error
- **Error classification**: sorting identified translation errors into error categories
- **Error identification**: the process of recognizing errors in a translation
- **Evaluation**: review of a translation in order to make a summative decision as to the quality of the translation
- **Fluency**: the quality of the source or target as a text on its own. (i.e, is the text linguistically well-formed and understandable?)
- **Freelance translator**: a professional translator who works as an outside contractor and not as an employee of a translation company
- **GALA**: Globalization And Localization Association, International non-profit industry association for companies that provide localization, language and technology services. (www.gala-global.org)
- **Holistic**: considering the entire translation at one time

- **Issue**: As issue is a potential problem detected in content. (Note: The term issue as used in this document refers to any potential error detected in a text, even if it is determined not to be an error. For example, if an automated process finds that a term in the source does not appear to have been translated properly, it has identified an issue. If human examination confirms finds that the term was translated improperly, it is an error. However, human examination might also find that the issue was not an error because the linguistic structure in the translation dictated that the term be replaced by a pronoun, so the translation is correct. Since issues may be automatically detected or incorrectly identified, this document refers to issues in most contexts.)
- **LISA**: The Localization Industry Standards Association was a translation standards body that was very influential in the translation industry until its closure in 2011. It is well known for its leading role in the development of the Translation Memory eXchange (tmx) and TermBase eXchange (TBX) industry standards among others.
- **Localization**: the adaptation of content to fit a specific locale
- **Metric**: a selection of error categories combined with how the metric is used for the purpose of measuring translation quality against the translation specifications
- **MQM**: Multilingual Quality Metrics (www.qt21.eu/mqm-definition) a framework for creating customized translation quality assessment metrics
- **Multidimensional**: considering multiple classes of error categories during a translation quality assessment or evaluation
- **Overall quality score**: measurement of translation quality for the entire text
- **Practicality**: regarding the speed or cost of using and assessment metric
- **Parameter**: general characteristics of a translation, such as target audience, register, delivery date, etc.
- **Provider**: those who provide translation services
- **QTLaunchpad**: Quality Translation Launchpad, a consortium of various organizations, including the European Union, DFKI, and GALA, with the focus of promoting translation quality for the 21st century. (http://www.qt21.eu/launchpad/)
- **Reliability**: for the purposes of this study, the inter-rater consistency between experts and non-experts using MQM metrics

- **Requestor**: those who initiate the translation process
- **SAE J2450**: a translation quality assessment metric created by the Society of Automotive Engineering for assessing translations in the automotive industry
- **Scorecard**: the usable implementation of a metric
- **Severity**: An indication of the how severe a particular instance of an issue is. Issues with higher severity have more impact on perceived quality of the text. The default MQM severity model has three levels: minor, major, and critical.
- **Specification**: defined features of a particular translation derived from a generalized set of translation parameters
- **TAUS**: Translation Automation User Society (www.taus.net)
- **Transcreation**: a translation of a text which is not a word-for-word translation but is loosely based on a source text and highly adapted to the target audience and purpose.
- **Translation aspects**: who, what, where, when, why, and how an assessment takes place
- **Translation service**: tasks performed by the translation provider on behalf of the translation requestor
- **Translation stakeholders**: participants in the translations process, including requestors, providers , and end-users
- **Translation quality threshold**: acceptable quality score before rejecting the translation
- **Type I error**: the identification of an error that is not truly an error (false-positive)
- **Type II error**: the failure to identify an error
- **Validity**: for the purposes of this study, an MQM metric is valid if it has all needed error categories for the assessment
- **Viability**: for the purposes of this study, the reliability, validity, and practicality of MQM metrics
- **Weight:** A numerical indication of the how important a particular issue type is in overall quality assessment. The default weight for issues is 1.0. Higher numbers assign more importance to an issue type, while lower numbers assign a lower importance. A weight of 0 would indicate that an issue is checked but not counted in MQM scores. Weights serve as multipliers for error penalties in MQM scoring.

- **Workflow**: the sequence of tasks in a translation process, beginning from the initial source text and resulting in the final target text along with all associated communication and approvals

WORKS CITED

Drugan, J. (2013). Quality in professional translation. (1st ed.). London: Bloomsbury Academic.

Eckes, T. (2011). Introduction to many-facet rasch measurement. Frankfurt: Peter Lang.

GALA survey results: initial results given to all participates, final results to be published soon.

Garvin, David A. "What Does Product Quality Really Mean?" Sloan Management Review 26.1 (1984): 25-43.

House, J. (1997). Translation quality assessment: a model revisited. Tübingen: G. Narr.

Melby, A. (2012) Brigham Young University Human and Machine Translation Quality: Definable? Achievable? Desirable? (forthcoming)

Melby, Alan, Paul Fields, Daryl Hague, Geoffrey Koby, and Arle Lommel. "Defining the Landscape of Translation." *Tradumàtica* 1 Dec. 2014a: 392-403. Print.

Melby, Alan, Paul Fields, Daryl Hague, Geoffrey Koby, and Arle Lommel. "What is Quality? A Management Discipline and the Translation Industry Get Acquainted." *Tradumàtica* 1 Dec. 2014b: 404-412. Print.

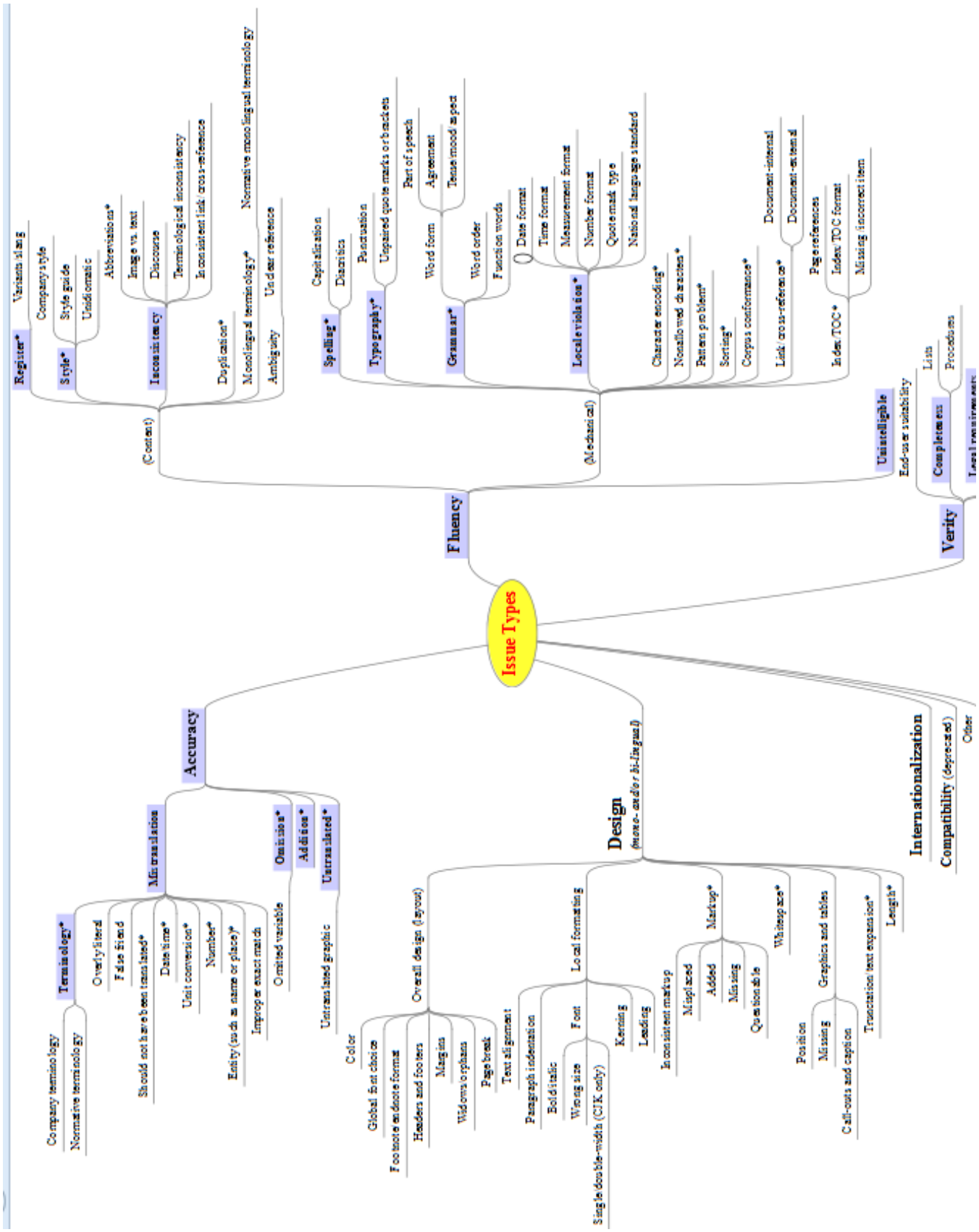
Melby, Alan, Paul Fields, Daryl Hague, Geoffrey Koby, and Arle Lommel. "Defining Translation Quality." *Tradumàtica* 1 Dec. 2014c: 413-420. Print.

O'Brien, S. (2012). Towards a dynamic quality evaluation model for translation. The Journal of Specialized Translation, (17), 55-77.

Williams, M. (2009). Translation quality assessment. Mutatis Mutandis, 8 (1), 3-23. Ottawa: University of Ottawa Press.

Wilson, M. (2005). Constructing measures: an item response modeling approach. Mahwah, N.J.:
Lawrence Erlbaum Associates.

Appendix A: Full MQM Hierarchy
(Online reference at http://qt21.eu/mqm-definition)



Appendix B: GALA-GLOBAL Stakeholder Survey 2013 Questions

1. In which country are you based?

2. Are you responsible for purchasing translation services?

3. What is the name of your organization?

4. What is your role in this organization?

5. How many employees does your organization have?

6. How many employees in your organization are involved in translation, localization, or globalization activities?

Academia/education/research

Government/NGO

Industrial user of translation services (client/end-user)

Provider of translation technology

Single-language vendor

Regional-language vendor

Multiple-language vendor

Translation/language service provider

7. Which of the following best describes your organization?

Academia/education/research
Government/NGO
Industrial user of translation services (client/end-user)
Provider of translation technology
Single-language vendor
Regional-language vendor
Multiple-language vendor
Translation/language service provider

8. What percentage of your translation content do you outsource to other organizations?

9. Which of the following describes the level of technology implementation in your organization?

We're cutting edge: We try the newest technologies, we use sophisticated technology that we developed in-house, etc
We are very technical: We use well-known tools for project management, CAT, terminology, etc.
We are somewhat technical: We use basic computer technology, but rely primarily on manual processes
We are non-technical: We don't use much technology but would like to
We are non-technical: We don't use much technology and do not see a need to
I don't know

10. Which of the following describes process development in your organization with regard to translation/localization tasks?

We have highly developed and defined processes that we use for everything
We have defined processes that we use for most tasks
We have informal processes that we use for most tasks
We do not have formal processes
I don't know

11. What are the primary source languages of content, in terms of volume, that you have translated?

12. What are the top target languages of content, in terms of volume, that you have translated?

13. Have you seen a trend towards translation into smaller or more exotic languages(e.g. languages for emerging markets) in your work?

14. For each of the following. Please rate how demand for translation services is changing in your organization for the following types of content:

Answer Options	Demand is increasing	Demand is steady	Demand is decreasing	I don't know
Marketing				
Technical Documentation				
Web content				
Online help				
Tutorials/E-learning				
Entertainment				
User-generated content				
Mobile Applications				

15. Please rank the vertical industry segments in which you translate or require translation

(1 =largest volume):

Automotive
Banking/finance
Consumer products
Entertainment/media/gaming
Government/NGO
Information Technology
Legal/IP/Patent
Medical/health care
Science/research
Telecommunications

16. Which quality assurance and/or assessment models do you currently use for human translation (Select all that apply)

ISO 9000 series
Six Sigma
EN 15038
SAE J2450
LISA QA Model
TAUS Dynamic Quality Framework (DQF)
Tool-specific model (e.g. whatever is implemented in your CAT or QA tool)
Internal model
We do not use a formal model for these tasks
I don't know

17. Which of the following quality assessment tools, if any, do you use? (Select all that apply)

Acrocheck
Built-in CAT tool functionality
ApSIC XBench
ErrorSpy
LISA QA Model Software
Okapi CheckMate
Yamagata QA Distiller
In-house tool
We do not use any QA tools

18. If your processes generate a quality score, for which of the following tasks do you use it? (Select all that apply)

Determining pricing
Acceptance testing for translation
Evaluating/selecting translators
Training translation staff
Risk management
Determining whether or not to publish content
Evaluating capability for addressing new markets
Not applicable

19. How often do you evaluate the quality of human translation? (Select one option)

Never
One-off checks when there are "changes" (e.g. new translator, new client, new domain)
Regular checks (according to some pre-defined criteria and/or schedule)
Random checks (no fixed criteria or schedule; when circumstances allow or require)
Always and systematically
I don't know

20. Do you use any sort of translation specification/brief to instruct LSPs/translators on your expectations and, if so, what sort? (Select all that apply)

We do not use any sort of translation

specification/brief
We provide an informal description of the project to translation staff
We use a formal specification/brief based on ASTM F2575
We use a formal specification/brief based on ISO/TS 11669
We use formal specifications based on our own in-house system
I don't know

21. Do you see the need to change your quality assessment processes in the next 2 years?

22. Which of the following best describes your use of machine translation (MT) to meet your translation requirements?

We currently use MT
We do not currently use MT, and we have no plans to use it
We do not currently use MT, but we are planning on using it within the next year
We do not currently use MT, but we are planning on using it in the future (but not within the next year)

23. For what percentage of your outbound translation requirements (i.e. translation of your content for consumption by others) do you currently use machine translation?

24. Are you seeing increased requirements for high-quality machine translation (versus "gist"/"information only" quality) in your work?

25. What sorts of machine translation systems do you use? (Select all that apply)

Rule-based MT
Statistical MT
Example-based MT
Hybrid MT
External online services such as Google, Babelfish, or Bing
I don't know

26. In which of the following areas have you customized some or all of your MT systems in your workflows? (Select all that apply)

Terminology
Additional corpora
Linguistic rules (e.g. RegEx)
Controlled authoring/language
Implementing data from output back into the system(s)
Improving the quality of existing data (e.g. corpora, rules)
We have not customized our MT systems
I don't know

27. How do you determine if MT meets your quality needs? (Select all that apply)

We use human evaluation (e.g. internal staff or external experts assess fluency and adequacy, etc.)
We use automatic scoring with standard metrics such as BLEU, METEOR, TER, etc.
We use automatic scoring with in-house / internally developed methods
We use both human and automatic evaluation methods as listed immediately above
We don't assess MT quality
I don't know

28. In your opinion, how good is the output quality of the machine translation system(s) you use?

29. How much of your outbound machine translated content do you post-edit?

30. Who carries out the post-editing?

In-house translators
In-house post-editors
Other in-house roles (e.d. editor, QA, SME)
External translator
External post-editor
I don't know

31. How do you assess the quality of post-edited content?

We use human evaluation (e.g. internal staff or external experts assess fluency and adequacy, etc.)

We use human evaluation (e.g. internal staff or external experts assess fluency and adequacy, etc.)
We use automatic scoring with standard metrics such as BLEU, METEOR, TER, etc.
We use automatic scoring with in-house / internally developed methods
We use both human and automatic evaluation methods as listed immediately above
We don't assess quality for post-edited content
I don't know

32. Do you have any additional comments on the subject of this survey that you would like to share?

33. Would you be interested in participating in further research on translation quality and technology usage?

34. If you would like to receive a detailed copy of the survey report? Please provide the name and e-mail address to which you would like to have it.

Appendix C: Firefox Specifications

Section 1: Language/locale (source) and (target)

- Source Language: en-US
- Target Language: es-MX
- The target language DOES pose particular grammatical or stylistic difficulties
- The target language DOES NOT use a different writing system than the source
- The author WAS a native writer of the source language
- It is unknown whether the translator was a native speaker of the target language
- The source text IS NOT already a translation of another text (e.g., a pivot language translation)

Notes: For question 4, most will be native speakers, for some, Spanish is their second language.

Section 2: Subject field/domain

- Subject Field(s): Software/Internet
- The subject field IS NOT in a regulated industry where legal compliance is mandated
- The subject field IS a technical field where particular terminology is expected

Notes: User interface for desktop Firefox browser.

Section 3: Terminology (source/target)

- Terminology: Terminology project within mozilla.locamotion.org/es_MX/terminology in PO format.

Notes:

Section 4: Text type

- Text Type): User interface elements
- The text type IS likely to require special attention to style or accuracy
- The text type WILL have other implications for other aspects of the translation

Notes: Some code will be present, like accesskeys, variables, CSS rules, blank space, tailing and leading white space. These are .dtd and .properties files, so translatable values will be assigned to entities that are referenced within the main source code.

Section 5: Audience

- Audience): Internet users
- The audience DOES require particular consideration in terms of reading levels, terminology, or style

Notes: Terminology, see section 3. Style, Mozilla has a particular style for content within the Firefox browser that people are accustomed to, which is typically a lower register. Audience will range from kids, to adults with various levels of education and background in Mexico. Some may even have Spanish as their second language. This localization is also used as a pivotal localization for indigenous Mexican localizations of the Firefox browser.

Section 6: Purpose

- Purpose: Provide accessibility to the internet for Spanish-speaking users in Mexico and Central America.

Notes:

Section 7: Register

- Register: informal
- There ARE conventions or expectations regarding register in the target language for this text

Notes: El tuteo.

Section 8: Style

- Style: Mozilla corporate style guide as well as Mexican l10n team style guide
- There IS a formal style guide that should be used for the text

Notes: https://www.mozilla.org/en-US/styleguide/communications/translation/ https://developer.mozilla.org/en-US/docs/L10n_Style_Guide

---

Section 9: Content correspondence

- Content Correspondence: fullCovert
- There ARE reasons why problems with the source should be preserved in the target
- Claims about pricing, availability, or other aspects of business WILL APPLY to the likely locale of the end-user.
- A Summary translation IS acceptable

Notes: Anything that has the potential to break the build by changing it between source and target will need to be preserved in the target. Summary translation is only acceptable if the actual translated string proves to extend beyond it`s alloted UI space.

Section 10: Output modality

- Output modality: other

- The output modality DOES create constraints on the translation in terms of size, time, character set, legibility, or other areas that might impact how the text can be used

Notes: Per string size restrictions, and even byte size limitations. Character set needs to support extended Latin-based characters.

Section 11: File format

- File Format: .dtd, .properties, .lang
- The file format DOES NOT allow for styled (rich) text
- There IS a particular layout expected in this file format

Notes: Must reflect the source structure.

Section 12: Production technology

- Production Technology: Translation Memory + Terminology Management

Notes: Localization performed on Mozilla instance of Pootle.
http://mozilla.locamotion.org/es_MX/firefox/