

# TRANSLATION QUALITY MEASUREMENT IN PRACTICE

**Riccardo Schiaffino**  
*Aliquantum*

**Franco Zearo**  
*Lionbridge Technologies*

**Abstract:** This paper provides an overview of the Translation Quality Index (TQI), a measurement methodology that can be used as a reliable indicator of translation quality. The authors have been developing the TQI methodology for the past five years. The TQI was first implemented for commercial use in 2004.

## 1. TRANSLATION QUALITY MEASUREMENT

### 1.1 How did you start working on translation quality measurement?

We have known each other for years—we both graduated in translation from the same university and currently live in the same area—but we had never worked together. In 2000, we were both giving presentations at the ATA Conference in Orlando, Florida. We started to talk about how people generally assumed that translation quality could not be measured. One thing led to another, and we decided to see if we could find a way to measure the quality of translation as a basis for process improvement.

In the following three years, we developed our research and gave three presentations on how to measure and control translation quality. At the same time, we applied our work at the companies for which we worked: Lionbridge in Franco's case and J.D. Edwards in Riccardo's. Riccardo's career at J.D. Edwards ended on December 31, 2003, following PeopleSoft's acquisition. However, he continued researching translation quality measurement, developing tools to help the assessment and measurement of translation quality, and developing what we called the Translation Quality Index, or TQI. In the meantime, Franco proceeded to implement the translation quality measurement methodology at Lionbridge. The spreadsheet that Lionbridge uses to measure translation quality is the result of our joint collaboration.

### 1.2 Why try to measure translation quality?

Some people say, "You cannot manage what you cannot measure." Applied to translation, this means that without some means to assess the quality of translation, it is not possible to improve translation quality, nor is it possible to know if the translation quality is good; and, if it is good, how to keep it that way.

### 1.3 Is it possible to measure translation quality?

We believe that it is possible to measure translation quality, although perhaps not directly: When measuring translation quality, we really measure the incidence of various types of errors and defects in the translated material; for example, errors of terminology, grammar, spelling, meaning, and others. Therefore, a good translation is one in which fewer errors are made.

Experienced translators would summarize the criteria for recognizing a poor translation as follows: “I know it when I see it” (Note 1). However, this simplistic approach is not adequate in meeting the demands of today’s high-paced business environment. Like many business processes where the desired outcome is a product or service, quality measurements are not only possible, but necessary. Without objective ways to measure the quality of our work, we are left at the mercy of fickle evaluations by lay people who can be highly subjective and not entirely fair.

We believe it is the translation profession’s responsibility to develop criteria that constitute an objective and fair evaluation of translation quality. Having said that, we heed the warning of the ATA. “Although the use of points may impart a certain impression of objectivity, it is in truth still subjective” (Doyle, 2003).

#### **1.4 Why measure errors when measuring translation quality?**

One important thing to consider is that the assessment of translation quality should be as objective as possible. What I like and what you like may be very different, but we should have some means to agree on certain standards.

We believe it is easier to agree on what constitutes an error rather than on what constitutes “quality” in the abstract, and that an important factor in quality is the absence of errors.

We also believe that summarizing all of the error points in a single index value will help us to synthesize the translation quality of a given text. Moreover, we can use statistical methods to determine if a translation process is in statistical control, if special causes are present, or even if we are improving our translation process.

#### **1.5 Do you believe there is one “ideal” translation process to ensure the best possible quality level?**

The process does not really matter as long as it yields the desired result. We believe that the very purpose of translation measurement is to obtain useful information for benchmarking the relative merits of various translation processes.

The real question then is, “What is the most efficient process in terms of quality versus cost?” We believe that the ingredients of good quality translations are fairly reasonable, but very seldom found all together. These ingredients include the following:

- Good translators with a sound linguistic and specific technical background
- Detail-oriented editors and knowledgeable proofreaders
- Thorough terminology work up front
- Sufficient time to provide a good translation
- Meaningful feedback and support from the customer

## 1.6 How does translation quality measurement differ from other methods of translation quality assessment?

Over the past 30 years, many methods of evaluating translation quality have been developed and proposed. Malcom Williams (2004) classifies these methods into two categories: Quantitative-centered systems and argumentation-centered systems. Williams characterizes quantitative-centered methods by some method of error counting, while argumentation-centered methods take a more holistic approach. Each method has its advantages and disadvantages, which we cannot elaborate here. Suffice it to say, the advantage of the quantitative-centered methods is that they lend themselves to quantifying errors and, therefore, make measurements possible.

## 2. THE TRANSLATION QUALITY INDEX (TQI)

### 2.1 What is the TQI methodology?

The TQI methodology (along with similar initiatives such as the LISA QA Model and SAE J2450) is a quantitative-based method of translation quality assessment. It measures the number and type of errors found in a text and calculates a score, or TQI, which is indicative of the quality of a given translation.

The distinctive traits of the TQI methodology are as follows:

- **Translation Quality Index.** The Translation Quality Index is a number that is indicative of the quality of a given translation. It is obtained by the rigorous application of a quality assurance methodology.

The Translation Quality Index attributes a value to a translated text, with 100 being an “error-free” translation. It is based on the number of error points in a given text or sample. Negative values are possible. The TQI is analogous to a temperature scale. We all have subjective interpretations of “cold,” “warm,” and “hot.” The use of a temperature scale (Fahrenheit, Celsius, or Kelvin) makes it possible to move from subjective perceptions to objective measurements.

- **Separation between error type and severity.** There are no pre-assigned penalties for the different error categories. Each error can be marked as critical, major, or minor, depending on its consequences. Sometimes, an error can be classified in different ways; for example, if I type “car” instead of “cab”, it could be classified as a mistranslation, a terminology error, or even a typo. While a precise classification of translation errors might be of interest in an academic setting, such as translation training programs, it is often unnecessary in a business environment.
- **Strict criteria for the severity levels of errors.** A TQI measurement should be objective, reproducible, and repeatable. To achieve these criteria, the evaluator has to follow certain rules when marking errors.

## **2.2 What are error points and how do error points differ from errors?**

Using a typo as an example: if we find five typos, we count five errors. That is a rather simple form of error measurement, but not all errors are equal. There is a difference between a typo on the front cover of a manual and the same typo in a footnote. There are also typos that alter the meaning of a word, and typos that do not lead to confusion; for example, the word “\*attention” spelled with three ‘t’s. This observation prompts us to assign different weights to errors depending on their consequences. In our previous example, we can decide to give minor typos a weight of “1,” and major typos a weight of “5,” “100,” or whatever. We call these weights “error points.”

## **2.3 What were the difficulties when you started to put the TQI into practice?**

The purpose of the TQI and its ancillary tools is to make translation assessment as objective as possible. However, when we started to use the TQI tool, we realized that how we configured the score was not always a true representation of the translation quality. It is easy to form an idea about how good or bad a translation is and then semiconsciously try to convince oneself that a major error is minor, or a minor error is only a “preference,” so as not to push the TQI below the threshold that would make the translation fail. Also, accuracy errors are difficult to evaluate when there is a slight loss in meaning. Even grammatical errors are sometimes not as straightforward as one would think. Language, after all, is not a precise science.

## **2.4 What makes a good evaluator?**

A good evaluator must be able to be as objective as possible. He or she must be able to distinguish between factual, tangible errors and stylistic preferences. We all have our pet peeves when it comes to translation choices. An objective evaluator realizes that he or she might have translated a sentence differently, but that the version chosen by the original translator is also acceptable.

You can roughly classify evaluators into purists and descriptivists. The purists are those who like to think of language in terms of how it ought to be used. Descriptivists, on the other hand, take into account how people use the language in their daily lives. Each point of view has its pros and cons, and they each lead to very different interpretations of what is considered “right” and “wrong.”

Moreover, if you give the same translation to two different evaluators, chances are that they will find a different number of errors or mark the same errors differently. A better solution would be to have the translation evaluated by a group of evaluators, in the same way that gymnastics resort to a panel of judges. Unfortunately, this solution proves to be too expensive in most commercial settings.

What would be helpful is a certification program for evaluators, possibly sponsored by an independent, not-for-profit organization such as the ATA. This not-for-profit organization might create standards regarding error classification, severity levels, error points, and others.

## 2.5 How do you distinguish between errors and stylistic preferences?

Bruno Osimo says that translation is a process with one entry point and multiple exit points. (2004). As discussed earlier, there is more than one way to translate a given sentence, each version being roughly equivalent and any differences being a matter of style and personal preference.

By definition, stylistic preferences are not errors and are ignored in the computation of the quality score. Therefore, it is necessary to establish clear rules that define what is an error and what is not an error.

We have developed a three-pronged rule to determine whether a marked error is preferential or not. Basically, the evaluator has to answer the following three questions:

1. Is it grammatically correct?
2. Is the translation accurate?
3. Is the translation compliant with the glossary, style guide, guidelines, and client instructions?

Answering the first two questions is not as easy as it might seem. In the case of grammatical correctness, for example, some languages might have authoritative language bodies; for example, *Real Academia de la Lengua Española* in Spain; *Académie française* in France; *Nederlandse Taalunie* in The Netherlands, and so forth. Other languages that lack such language authorities, such as American English, might have to rely on commonly accepted language conventions as described in authoritative reference books; for example Merriam-Webster's Dictionary, *The Chicago Manual of Style*, and others. A third group of languages does not have established language conventions, as is the case with many languages in India. In such cases, it is important to develop glossaries and style manuals.

Evaluating the degree of accuracy is another challenging task. We have developed flow charts similar to those created by the ATA for test evaluation for certification purposes. The intent is to see if there have been significant deviations in meaning.

The last question serves the business purpose of delivering quality that conforms to the client's specifications. This is generally more straightforward: Either the term is in the translation glossary or it is not. Either the translators followed the style guide and the instructions, or they did not.

## 2.6 Can the TQI help in assessing the quality of machine translation?

Absolutely. Some argue that human-based evaluations are too subjective, that MT should not be evaluated using human-based methods, and that such evaluations are too subjective. We do not agree. The TQI is a sort of Turing test. The Turing test was developed to indicate whether a machine was "intelligent" by testing its capability to perform human-like conversation. If a user cannot tell the difference between a text translated by a human and one by MT, then we could say that the two texts are equivalent. The TQI can help with this evaluation. If we agree that a TQI score of 80 or above is the mark of a good translation, it does not matter which localization process we used to obtain the score. In our experience, raw MT outputs have TQI scores below

zero. Processes that combine MT with human post-editing can elevate the TQI scores to levels that are more acceptable.

## **2.7 Is there anything that the TQI methodology cannot measure?**

Yes. In our experience, there are a couple of cases where relying on the TQI methodology would be inappropriate.

Because the TQI methodology is designed to measure tangible, factual errors, it shows its shortcomings when it comes to evaluating so-called “literal” or “word-for-word” translations. A literal translation might comply with the three-point preferential rule, grammaticality, accuracy, and compliance, and still be regarded as a poor translation.

Another case where the TQI methodology proves to be ineffective is when a high degree of creativity is expected on the translator’s part, which is often the case with translations for marketing and advertising. In these types of text, translators and copyeditors might have a certain degree of freedom. It is an acceptable practice to deviate from the source text as long as the translator maintains the core message. Conversely, the TQI system penalizes deviations from the source text as accuracy errors, something that a translator in other circumstances is not allowed to do.

In our experience with marketing texts, the translation might contribute 60-75% of the final version, the remainder coming from additions, deletions, and textual changes as deemed appropriate.

## **3 ADDITIONAL RESOURCES**

Copies of our previous presentations, a translation quality web blog, and other materials can be found on our website, [www.translationquality.com](http://www.translationquality.com).

## **4 NOTES**

1. We are referring to U.S. Supreme Court Justice Stewart’s remark about the difficulty in finding an objective definition for an obscene motion picture. In *JACOBELLIS v. OHIO*, 378 U.S. 184 (1964) he remarked:

“I shall not today attempt further to define the kinds of material I understand to be embraced within that shorthand description; and perhaps I could never succeed in intelligibly doing so. But I know it when I see it, and the motion picture involved in this case is not that.”

## **5 REFERENCES**

1. Doyle, Michael Scott (2003). “Translation Pedagogy and Assessment: Adopting ATA’s Framework for Standard ErrorMarking”, in *The ATA Chronicle*, November/December 2003.

2. Osimo, Bruno (2004). *Traduzione e qualità: la valutazione in ambito accademico e professionale*, Hoepli, Milano, p. 25
3. Williams, Malcom (2004). *Translation Quality Assessment: An Argumentation-Centred Approach*, University of Ottawa Press, Ottawa