A DEEPER LOOK INTO METRICS FOR TRANSLATION QUALITY ASSESSMENT (TQA): A CASE STUDY

ROBERTO MARTÍNEZ MATEO
Universidad de Castilla-La Mancha
roberto.martinez@uclm.es

1. Introduction to TQA

The search for quality in translation is still an unsettled issue today. From the second half of the 20th century onwards, controversy surrounding the quality concept and the way to determine it has become central. Nonetheless, it seems that there is no common ground when it comes to defining quality either from a practical or from a theoretical viewpoint. Moreover, there are many scholars who still believe that quality in translation is a *relative* and *subjective* concept (Horguelin and Brunette 1998; Larose 1998; Parra 2005). Not surprisingly, it has been the excess of conflicting opinions and the experts' lack of consensus on the definition of quality that have hampered any potential progress in the field (Colina 2009).

However, there seems to be a general agreement on some points, such as what are the optimal measures to be taken when building a successful model. In order to assess quality in translation three steps should be taken: firstly, quality must be defined. Many would agree that a quality translation is one which fits its purpose (Nord 1997; O'Brien 2012). Secondly, the methodology must be set. For that, special attention has to be paid to those quality assessment methods that enable measurement[1]. And thirdly, the assessment should be carried out in accordance with the definition of quality as applied to the text and to the assessment methodology chosen.

2. Definition of Quality

Different authors offer various definitions of translation and quality and of translation quality, which are basic concepts of any translation theory. These notions are so ample that different translation theories may put forward their own view (Gómez 2002). Subsequently, different views of translation give shape to different concepts of translation quality and so call for different ways of assessing it (House 1997).

Quality is far too complex a matter and too dependent on context (Nord 1997) to be condensed to an all-embracing definition. It has to do with a wealth of factors: fulfilling user needs or expectations, enhancing work efficiency, profitability, deadline compliance, resources and tools availability, etc. These characteristics (and many others) that could be attributable to quality do not all have the same weight on each translation assignment and are not therefore equally measurable or assessable.

A review of quality evaluation literature from industrial sectors has revealed that most quality standards define the concept as the ability to fulfil a client-defined set of parameters (Jiménez-Crespo 2009)[2]. Nonetheless, in translation, the concept of quality has traditionally been linked to values such as accuracy, correctness and fidelity (to the original). Currently, the concept has evolved to take on a higher polyhedricity due to the fact that quality can be observed from diverse angles[3] and, thus, checked at different stages and with regard to objects. Therefore, delimiting this intricate concept calls for the assumption of a multifaceted view.

It seems reasonable to think that given the subjectivity and relativity of the notion, and indeed of the evaluator (House 1997), quality assessment requires something that could offer the process greater objectivity. Without explicit criteria on which to base evaluation, the evaluator can only rely on his/her own view (Colina 2009). As a result, fixing a number of parameters or criteria as a yardstick for comparing real versus ideal performance could remove a great part (but not all) of the subjectivity and could lead to a higher inter-rater reliability (Doyle 2003; Colina 2008, 2009).

The view of (translation) quality in this paper is equated with the notion of *adequacy*, in the functional sense, considering quality to be the appropriateness of a translated text to fulfil a communicative purpose. It is thus a dynamic concept related to the process of translational action (Nord 1997). Hence, TQA methods have to be flexible and customizable enough to cater for as many scenarios as possible. For that, a comprehensive measurement procedure that incorporates a holistic evaluation (Jiménez-Crespo 2011) would be required.

3. TQA: the What

Colina (2009: 236) states that TQA "is probably one of the most controversial, intensely debated topics in translation scholarship and practice". The bibliographical review has revealed that both the concept and the terminology of the field overlap (Conde 2008). Nonetheless, the process of evaluating translation quality is widely known as Translation Quality Assessment (TQA) (Parra 2005). Many proposals for TQA have already been laid on the table, but none of them has proved to be a definite solution. What is more, the search for a unique method for TQA that could achieve full objectivity in every situation, context and for every type of text seems illusory.

Waddington (2000) warns that the object of assessment must be specified in order to avoid misunderstandings and to carry out a valid assessment. According to Stejskal (2006: 13), quality in translation can be analysed in what he calls the "3Ps" of quality assessment: producer, process and product. The procedures, measures, tools for evaluating quality in each of these instances have nothing to do with each other and, besides, focus on different dimensions. In this case, evaluation focuses on the Product adopting a textual approach (House 1997) to value the linguistic quality of the output.

The quality of the producer can only be evaluated by means of certification and, as Stejskal (2006: 13) points out, this "occurs under three possible scenarios: certification by a professional association, certification by a government, and certification by an academic institution".

As for the process**,** standards have become their measuring rod. They are process and not product-oriented (Martínez and Hurtado 2001) and their basic tenet is that when predefined processes are followed, good results (translations) will be obtained. In Europe, for example, the CEN (*Comité Européen de Normalisation*) approved in 2006 the EN15038 standard whose main aim is "to establish and define the requirements for the provision of quality services by translation service providers". Nonetheless there is not yet available an international standard exclusively designed for translation and some scholars forecast that there never will be. As Secâra (2005: 39) remarks, "The reason why no single standard will suffice is that quality is context dependent". As a result, current TQA tendencies have opted for a more restrictive view by focusing on the product.

On the whole, the product-centred methods are divided into two branches. One of the trends examines the linguistic features of translated texts at sentence level, that is to say, using an error-based translation evaluation system as the procedure for quantifying quality (Secâra 2005), whereas the other trend highlights macrostructure relations of the text as a unit. Waddington (2000) calls the first

type *quantitative-centred* (bottom-up) systems and the second the *qualitative-centred* (top-down) systems. Roughly speaking, this is what Colina (2009: 237) calls *experiential* and *theoretical* approaches, respectively. According to Williams (2004) the first type includes the *quantitative-centred* (error counting) systems and the second the *argumentation-centred* (holistic) systems.

This article analyzes some quantitative systems for TQA, the so-called "metrics". Based on a typology of errors with a point-deduction scheme (depending on error type and severity), these systems count up these points and then subtract their negative value from the previously allocated bonus points. This operation gives a score that classifies the translation on a quality scale. Despite the drawbacks of metrics, as pointed out below, these quantifying systems fill in a gap in professional TQA arena (Jiménez-Crespo 2011), where translation becomes a business with time (De Rooze 2003) and budget (O'Brien 2012) constraints and so deserves to be studied.

4. Metrics for TQA: the How

Henceforth, various quantitative-oriented models for TQA are analyzed. This review includes the SICAL, the LISA QA model, the SAE J2450, the Quality Assessment Tool (QAT) and the TAUS Dynamic Quality Evaluation Model. Special attention is paid to a prototype tool developed by the Directorate General for Translation (DGT) of the European Commission as an aid in the quality quantification process of external translations.

4.1. SICAL

In the 70s, the first steps towards creating a more systematic and objective model for professional TQA were taken within the Canadian government's Translation Bureau with the creation of SICAL (*Système Canadien d'appréciation de la Qualité Linguistique*[4]). This system aimed at discarding the evaluator's value judgement traditionally dependent on his particular knowledge and appraisal (e.g. the translation is "accurate" and "reads well" or the "translator's choice is clumsy and vague", in Williams 1989: 14). SICAL I established a revision process at microlinguistic level that carried out a contrastive linguistic analysis of the pair of texts (ST and TT) based on an error typology. This system fixed a set of reference parameters with which to compare the linguistic features of finished translation and this is how the concept of acceptability threshold for a translation came up, the fixing of a borderline between the acceptance and rejection of a translation. The final result is obtained by dividing the aggregate negative points (errors) by the

number of words of the text (usually a 400-word passage). Later, the system evolved into two general error categories (transfer and language) and, subsequently, these were classified according to seriousness into *minor* and *major*. Both definitions stressed the term *essential* as the defining characteristic for setting the acceptability limit (Williams 1989). A translation might pass with "as many as 12 errors of transfer, provided no major error was detected" (Williams 2001: 330). Therefore, major errors had to be unequivocally recognisable by the prospective raters. A drawback of SICAL was that it could have as many as 675 errors (300 lexical and 375 syntactic), which made its application a cumbersome task (Martínez and Hurtado 2001). Another downside of this system, as Secâra (2005) points out, is that the sample reviewed (approx. 400 words) is chosen randomly, what may raise doubts about its representativeness. All in all, SICAL was a milestone in TQA and paved the way for future developments.

4.2. LISA QA Model

The LISA Quality Assurance (QA) Model was developed in 1995 and distributed by the *Localization Industry Standards Association* (LISA) for localization projects. It is a stand-alone tool applied to product documentation, help and user interface, and even to computer based training (CBT) (Parra 2005). Its user-friendly interface comprises a series of templates, forms and reports embedded in a database (Stejskal 2006). Besides, it contains a predefined list of error levels of seriousness and relevance, a record of error categories, a catalogue of the reviser's tasks and a template for marking the translation as Pass or Fail (acceptability threshold). However, the tool is flexible enough to admit customization and allows the translator to reach a prior agreement with the customer on two key parameters: error type and severity (Parra 2005).

The LISA QA model version 2.0 appeared in 1999 and accommodated upgraded capabilities: Linguistic Issues, Physical Issues, Business and Cultural Issues and Technical Issues (LISA 2007: 12-14). The third version (3.0), completely revised, came out in 2004 (Parra 2005: 277) and was meant: "to define and experiment with their own quality metrics".[5]

However, in spite of the benefits of this new version, Jiménez-Crespo (2009) claims that its error typology lacks an empirical base and some of the error categories overlap, such as accuracy and style. In addition, as Parra Galiano (2005) perspicaciously points out, the norm does not define clearly what a translation error (*mistranslation*) or a style error is.

The LISA QA model also established an application procedure consisting of several steps. One or several samples undergo Quality Assessment (QA) using a template. When the TT fails, the rater adds remarks. When the TT passes, the rater carries

out a full revision or a Quality Control (QC) and adds the amendments later. Once the revision and the correction tasks are over, the TT undergoes another QA (Parra 2005).

The LISA QA Model[6] is a componential model made up of eight items, out of which only one covers language matters (Jiménez-Crespo 2009). Within this linguistic item, the LISA QA Model distinguishes seven error types: *Mistranslation*, *Accuracy*, *Terminology*, *Language*, *Style, Country* and *Consistency* (Parra 2005: 280-281). Each error of this typology, in its turn, may have an effect on the TT and is consequently classified in three degrees of seriousness: minor, major and critical depending on whether the mistake is not important (1 point), whether the error is detected in a visible part of the document (5 points) or whether it is located in a preeminent part of the document or may cause a bug (critical), respectively. To be acceptable, the TT must contain no critical error and the ratio between error points and total words cannot surpass a set figure.

4.3. SAE J2450

A working group made up of SAE and GM representatives developed this metric system. It was first introduced as a *Surface Vehicle Recommended Practice* in 2001 and turned into a standard in 2005. It set out to "be regarded as only one element in a total Quality Assurance Process, albeit an important one" (SAE J2450 2001: 2). This statement reminds us that any metrics are just that, a link within the chain of actions whose aim is to guarantee, check and improve translation quality.

Initially, this tool was to be used for revising service automobile documentation so that it could provide a "consistent standard against which the (linguistic)[7] quality of the automotive service information can be objectively measured" (SAE J2450 2001: 1). Unlike this metrics, it was not intended for translations where characteristics such as style, register and tone might play an important role (marketing, advertising translations or the like). Its application, although with adaptations, has recently spread to other industrial sectors such as Biology (pharmaceutical, medical devices, etc.).

Regarding its scope, the SAE J2450 does not specify how to select the sample, or its size, nor any specific acceptability threshold or, it follows, any advice on what to do with the assessment findings either. The norm openly admits that it only deals with linguistic error detection, but leaves aside style and format features. In addition, it does not attempt to explain the causes of errors but just detects, tags and counts them. It is the only metrics that even counts as errors those brought about by errors in the original text, which the translator has faithfully conveyed

into the TT. The errors are classified first according to their type in one of the seven main categories ranked in order (Wrong Term, Syntactic Error, Omission, Word Structure or Agreement Error, Misspelling, Punctuation Error, Miscellaneous Error), and secondly, according to their severity in one of the two subcategories (minor and major). Errors have a fixed penalization schema that can be adapted, but only on a global basis for each project. The points allocated to each error type are shown in Figure 1:

Main Category	**(abb.)**	**Sub-Category** (abbreviation)	**Weight** serious-minor
Wrong Term	(WT)	serious (s)	5/2
Syntactic Error	(SF)	minor (m)	4/2
Omission	(OM)		4/2
Word Structure or Agreement Error	(SA)		4/2
Misspelling	(SP)		3/1
Punctuation Error	(PE)		2/1
Miscellaneous Error	(ME)		3/1

FIGURE 1: SAE J2450 Translation Quality Metric. *© SAE J2450, Committee*

Likewise, it is admitted that error classification "is necessarily[8] a judgement call by the evaluator" (SAE J2450 2001: 3). The metric system aims at limiting the unavoidable subjective burden of the reviser by providing him with a reference error typology easy to apply accompanied by two metarules. These metarules (SAE J2450 2001: 4) are to be applied by the reviser in case of doubt:

1) *when in doubt, always choose the earliest primary category; and*

2) *when in doubt, always choose 'serious' over 'minor.'*

Their goal is to guide evaluators when they come across a dubious classification of errors. These metarules notwithstanding, the norm openly acknowledges the arbitrariness of setting these two metarules, while it argues that the consistent application of these metarules favours systematicity in the evaluators' decision-making process and, therefore, promotes reproducibility and repeatability.

The norm also stipulates a review process for the rater that consists of five steps in chronological order: 1) mark the error in TT (also repetitions), 2) choose the primary error category, 3) choose the secondary error category, 4) deduct the points and 5) calculate the final mark dividing the aggregate points by the number of words of the text.

4.4. The Quality Assessment Tool (QAT)

This QAT was developed within the Directorate-General for Translation (DGT)[9] to help revisers with external translation assessments. It belongs to the quantitative, "bottom-up" or experiential approaches to TQA. Internally, it is known as the "calculator" in reference to its main function, a computer-aided tool to quantify errors.

An internal audit carried out in 2008 by the IAS (Internal Audit Service) of the EC concluded that there existed diverging practices amongst different Language Departments (LD) regarding their freelance assessment approach. Therefore, it was recommended that DGT should endeavour to base external or freelance translation[10] assessment on quantifiable data as much as possible.

This tool was not devised from scratch. It took on the error typology that was used in the Translation Centre[11] (CdT). This typology included 8 error types: Sense (SENS), Omission (OM), Terminology (TERM), Reference Documents (RD), Grammar (GR), Spelling (SP), Punctuation (PT) and Clarity (CL), two error gravities (minor and major) and quality marking ranges. The QAT inserted slight modifications in relation to the CdT´s marking ranges as can be seen in the following table:

Mark	CdT (0-10 pt.)	QAT (0 -100 %)
Unacceptable	0-39	0 – 39
Below standard	40-59	40 – 59
Acceptable	60-79	60 – 69
Good	80-99	70 – 85

TABLE 1: QAT's mark ranges. *© European Union, 2013*

As with the previous tools, penalizing points are assigned to errors according to their type and seriousness. The final mark is obtained by deducting the aggregate penalizing points from 100% of the initial bonus. As a result, the translation is categorised within a quality range (vid. Table 1). The QAT also adopted the size of the assessment sample from the CdT, about 10% of the text, with a minimum of 2 and a maximum of 10 pages.

This Figure shows the interface of the QAT. The rater locates the file using the drop-down menu *Name* of file; next he chooses *Language* and the *Profile* to start working. There are three profiles available: *General, Political* and *Technical*. This choice exerts a great influence in the final mark since different profiles have

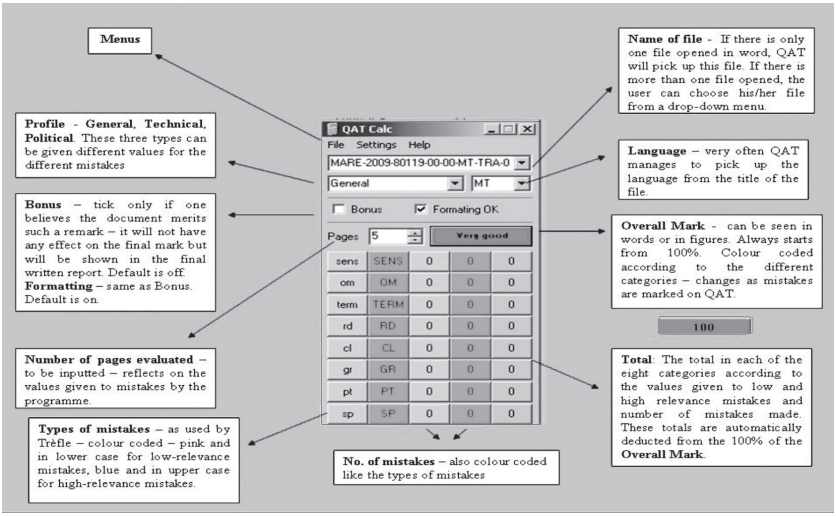

FIGURE 2: QAT's interface. *© European Union, 2013*

different deduction points for each mistake type and degree of seriousness. Next, the rater inserts errors by clicking on the corresponding button. Uppercase codes are for major relevance errors and lower case for minor ones.

All the parameters of this cluster are customizable, although some of them have preset values. The default profile is *General.* Besides, the default number of pages is 5 (considering a page to be 1500 characters with no spaces) and the number of initial credit points is 100 per 5 pages (20 points per page).

This tool allows multi-users. Unlike the SAE, the QAT only counts repetitions of the same error as one error. Below the Profile menu we find *Bonus* and *Formatting OK* checkboxes. By default, the *Bonus* is checked and *Formatting OK* unchecked. The reviser can tick them on or off but just for the sake of indicating that the translator has made good choices in his wording (*Bonus on*) or that the translation is neatly formatted (*Formatting OK*). The rater can activate the option *Comments* in the menu *Settings* to add any information he may deem fit to the errors that have been marked.

After completing the evaluation, the reviser can generate a report (Write report), which summarizes all the information about the revised text. This information includes in a .txt file the title, profile, source language, errors pinpointed, the weightings per error, as well as the final mark of the translation.

A prototype of the tool was presented by mid 2009; all DGT's Language Departments embarked on a trial phase during the second half of the year. The findings of this trial were gathered in 2010 and yielded varied conclusions depending on the LD. As a result, nowadays its application on an LD-basis remains optional.

The tool's greatest contribution is also its greatest weakness: the text type classification. This is the first quantitative TQA tool that contains text types (*Profile*). This initial choice conditions the whole assessment process. Each profile has a different set of weightings assigned to each error according to type and severity. The rater lacks sound criteria for classification. His pick of the Profile, in the absence of instructions, seems to rest on terminology density rather than on other clear parameters. Furthermore, the text classification as 'general', 'political', 'technical' or 'legislative' does not imply that neither a specific translation technique nor a particular evaluation method should be used.

Finally, the QAT stipulates that error repetitions will only be counted once. The question occasionally arises as to what a rater should do when, for instance, he comes across a variety of terms for the same concept.

4.5. TAUS Dynamic Quality Evaluation Model

One of the latest and most significant contributions to TQA has been the Dynamic Quality Evaluation Model. Sharon O'Brien (2012) developed this TQA model in collaboration with the Translation Automation User Society (TAUS).

Echoing the widespread feeling amongst its members that methods whose final score is simply based on counting the number of errors is too static and normative a system (O'Brien 2012: 55), TAUS carried out in 2011 a benchmarking exercise of eleven TQA models. Out of eleven, ten evaluation models were quantitative approaches (which included LISA QA model v. 3.1 and SAE J2450) and the remaining one was process-oriented.

An interesting finding of the benchmarking study was that error-based models sought to identify, classify, allocate severity level and apply penalty points to errors. They all have a pass/fail threshold and their analysis is made at the segment level, ignoring thus the larger unit of text (O'Brien 2012). Broadly, it is interesting to note that all the macro error categories identified coincide (including those of the process-oriented model). The most frequent macro error categories, and the micro error categories included in each macro, are listed in the following table.

This table shows that the prevalent error types are *Language*, *Terminology*, *Accuracy* and *Style*, despite their differing scopes.

Errors present in TQA models	Macro error type	Includes the following micro error types
10/11	Language	9/10 including grammar 7/10 including syntax 7/10 including spelling 6/10 including punctuation
10/11	Terminology	General consensus on definition: 1) Adherence to client glossary 2) Adherence to industry terminology 3) Consistency
9/11	Accuracy	7/10 including omissions 7/10 including additions 7/10 including inaccurate cross-references 7/10 including meaning
7/11	Style	4/7 including lack of adherence to 'client style guide'

TABLE 2: Macro and micro error types (adapted from O'Brien 2012: 60)

The benchmarking exercise reviewed some other evaluation procedures from related professional contexts (Machine Translation, Translation Training, Community Translation and Technical Translation). This review allowed us to shortlist the most common evaluation models, classified according to their control level from the most to the least (O'Brien 2012: 67):

(1) *Adherence to regulatory instruments*
(2) *Usability evaluation*
(3) *Error typology*
(4) *Adequacy/Fluency*
(5) *Community-based evaluation*
(6) *Readability evaluation*
(7) *Content sentiment rating (thumbs up/down, rating allocation)*
(8) *Customer feedback (Sales, Tech Support Calls etc.*)

This summary concluded that apart from the widespread error-counting systems, various other methods for TQA could be applied to professional translation. Consequently, a new proposal was put forward: the Dynamic Quality Evaluation Model. This model is based on two building blocks (Communication channel and content profile) and three evaluation parameters (Utility, Time and Sentiment[12]). In order to identify the various evaluation parameters used in professional

translation a company profiling survey was carried out amongst the participants. As a result, a number of different text types were mapped with different evaluation parameters. Thus, the Dynamic QE Model materialized in the following template (O'Brien 2012: 73):

Communication Channel	Content Profile	Uts Ratings	Recommended QE Models In Descending Order Of Control
		Utility: Time: Sentiment:	

Table 3: Dynamic QE model template (O'Brien 2012: 73)

According to the data gathered in the survey referring to the Communication Channel, the Content and the feedback on the parameters of Utility, Time and Sentiment, an ordered list of some evaluation models for TQA was proposed for five instances.

The main advantage of the DQE model is its adaptability to client preferences in terms of the quality parameters identified (UTS). Based on the type of content (eight parameters were identified)[13] and the communication channel used by the client (three were identified)[14], the model offers a shortlist of evaluation models in order of application. Therefore, it provides a customizable modular TQA system for the selected content types and quality criteria.

On the other hand, all the TQA models recommended for each instance are models with their own advantages and disadvantages, as is the case of error typologies. For example, the DQE model handles eight types of content but misses others, such as technical, legal, economic texts, to name but a few.

5. TQA Metrics Overview: Pros and Cons

Based on the foregoing review, the features of the quantitative-oriented models analyzed in the foregoing review will now be outlined. Moreover, a critical examination of these features has made it possible to list their advantages and disadvantages with a view to building a theoretical construction for a new TQA model.

It is observed that all these tools were created to be stand-alone and not plugged-into applications. They all apply Quality Control procedures[15] (Parra 2005)

(except for the SAE J2450 that also allows for Quality Assurance) highlighting their consideration for deadlines and resource-investment, two key factors in professional translation. Mostly, they take random samples to carry out a linguistic comparative analysis between the target and the source texts. Three of the systems (SICAL, LISA and QAT) coincide in setting the sample length at approximately 10% of the text. In all cases, the evaluation fulfils a summative function (Melis and Hurtado 2001) and is used to determine the end results and to judge whether the objectives have been achieved (criterion-referenced).

5.1. Weaknesses common to TQA metrics analysed

Next, the main weaknesses of quantitative systems are listed:

1. Firstly and most importantly, all the TQA metrics rely on rating scales that lack an explicit theoretical base and verifiable empirical evidence, as several scholars warn (Colina 2008, 2009; Jiménez Crespo 2001). This underlying theoretical defect results in a two-fold inadequacy: first, it damages their value due to their lack of a conceptual background and, second, it prevents these models from being revisited to be applied to other contexts or text types different from the originals.
2. Secondly, all the models analyzed here rely on the central concept of error as the defining element of their assessment model and, subsequently, of the related issues such as the error type, and severity and error weightings. As Parra (2005) stresses, some error categories are ill defined and some even overlap. She gives the LISA QA model as an example of an unclear definition of mistranslation or style errors. Hence, all these proposals shape their definition of a quality translation as an error-free text or a text whose number of errors (their allocated points) does not surpass the predefined limit (acceptability threshold). Moreover, all the proposals analyzed consider error as an absolute notion, disregarding its functional value (Hurtado 2001). Therefore, errors are identified and tagged in isolation and not in relation to their context and function within the text (Nord 1997). Furthermore, once an error is detected the problem is how to categorize it correctly within a type and a severity level. The red line that separates those categories is sometimes so thin or blurred that errors might be classified into different categories by different revisers.
3. These systems take care of linguistic related issues, but at a micro textual level, and pay no attention to textual or extralinguistic matters. Therefore, the search for errors is limited to the word and sentence tier and does not take heed of the larger unit of the text nor of the communicative context (Nord 1997; Williams 2001; Colina 2008, 2009).
4. In order to implement the assessment, the reviser carries out a partial revision (Parra 2007) of the selected sample. It seems reasonable, therefore, to question

the representativeness of the limited, variable-length sample (Larose 1998; Gouadec 1981). Besides, these metrics do not specify what type of revision has to be made (unilingual, comparative; vid. Parra 2005). Therefore, the subjectivity inherent to all human activity cannot be detached from these models since it is a person (reviser/rater) who has the final word in error detection and tagging.

5.2. Strengths common to TQA metrics analysed

Bottom-up approaches, despite not having been empirically tested, yield the following theoretical advantages:

1. What at first sight might seem a reductionist and simplistic definition of quality (error-based) of a humanly produced output (full of nuances) could, on the other hand, be seen from the opposite end. If a translated text contains no or only a few errors of a particular type, for instance terminology, this entails that the terminology has been suitably conveyed into it. Therefore, some repetitive macroerror categories can be identified. The comparison of error categories of these quantitative models is summarized in the following table:

METRICS	LISA QA Model	SAE J 2450	TAUS Benchmarking	QAT
ERROR TYPES		Miscellaneous		
	Accuracy	Omission	Accuracy	Omission
	Terminology	Wrong Term	Terminology	Terminology
	Language	Syntactic Punctuation Misspelling	Language	Grammar Punctuation Spelling
		Word structure or agreement error		
	Country		Country standards	
	Consistency		Consistency	Reference documents
	Style		Style	Clarity
	Mistranslation		Mistranslation	Sense
SEVERITY LEVELS	minor, major, critical	minor, major	minor, major, critical	minor, major

TABLE 4: List of error types and features of the quantitative models analyzed (SICAL is not included since its large number of error types makes it unmanageable)

It can be observed that most macro error types are consistently identified through time and the metrics. Some error types, then, are recurrently kept in all the models, although with term variation and slightly different scopes:

— Accuracy (LISA, TAUS)/Omission (SAE, QAT),
— Terminology (LISA, TAUS, QAT)/Wrong term (SAE),
— Language (LISA, TAUS)/Syntactic-Punctuation-Misspelling (SAE)/ Grammar-Punctuation-Spelling (QAT).

Three other error types are present in all the models, except for SAE:

— Consistency (LISA, TAUS)/Reference Documents (QAT),
— Style (LISA, TAUS)/Clarity (QAT) and
— Mistranslation (LISA, TAUS)/Sense (QAT).

However, some error types find no counterparts in other systems such as Miscellaneous (SAE) or only one, such as Country (LISA)/Country standards (TAUS).

2. These metrics present a clear quality categorization by setting an acceptability threshold and different quality ranges. Furthermore, their assessment relies on a predetermined error classification and transparent error weightings known a priori by all parties involved (Schäffner 1998). Since, after all, quality really boils down to an agreement between translator and "customer" on the kind of quality sought for a particular assignment.
3. As Hurtado (2004) points out, nowadays, when assessing translation quality in professional settings criteria such as return on investment cannot be omitted. In professional contexts, where time and resources are limited, TQA metrics are an efficient and timesaving proposal that offers a good value for money relationship and fills a gap in professional translation
4. Acknowledging that full objectivity in TQA seems to be a utopian aim, these metrics raise expectations of a high inter-rater reliability (Doyle 2003; Colina 2008, 2009) offering results that are valid, justified and defendable.
5. Metrics bestow systematicity and reproducibility on a process that necessarily requires human intervention (Hönig 1998).

6. Outlining a Model for TQA

The foregoing analysis aimed to develop a valid and reliable model for professional TQA that tries to remedy the deficiencies in the quantitative models highlighted in the analysis, with special attention to QAT. To this end, some fundamental changes will be made incorporating the positive contributions of qualitative

models. The new proposal has necessarily to bridge the existing gap between theoretical sophistication and the applicability (Colina 2008) of the existing models. To do so, it relies on the functionalist paradigm (Nord 2009) because it provides two main benefits: it offers (i) a sufficiently ample framework to solve the theoretical deficiencies posed in the models of the foregoing review and (ii) a pragmatic and textual approach that encompasses extratextual and pragmatic factors. Consequently, the *Fit-for-purpose* motto is taken as the main evaluation parameter and it is placed in a central position in the new proposal.

As the new TQA model is intended to be used in professional settings, the applicability required relies heavily on its use of easily understood, practical, limited in number and verifiable (Brunette 2000) quality criteria. But above all, these criteria have to be flexible and customizable to the specific situational context (Martínez and Montero 2010) able to assign the relative value of error.

Another important drawback of metric systems is that they do not duly pay attention to the contextual (Sager 1989), the pragmatic (Nord 1997) nor the text-level issues. To overcome these hurdles, this model takes a two-tier and a continuous methodological approach. At the first tier, at sentence level, and taking a bottom-up approach, the new model is grounded in an error typology based on the above identified dominant macro error categories (sense error, terminological error, reference documentation error, omission error, clarity error, spelling error, grammar error, punctuation error plus a new type, addition error). At the same time, this new tool adds a new classification of errors according to their nature. Thus, errors may be tagged as *pragmatic* (relative value) (Nord 1997; Jiménez-Crespo 2011), when the error becomes such in virtue of its context; or as *linguistic* (absolute value), when an item is deemed an error per se and is not context-dependent. This is a key distinction for comprehending and implementing the relative value that functionalism concedes to error. Accordingly, this theoretical stance considers error as an inadequacy in relation to their context and the goals it pursues (Nord 1997).

This first tier of analysis leads to the second one that takes place at text level from a top-down approach. Here, an assessment rubric (Moskal 2000) helps the rater to carry out a linguistic analysis from a holistic viewpoint (Waddington 2000). The rubric is an assessment tool that splits the object of study (quality concept) into smaller components (dimensions) to simplify its assessment. With the form of a double-entry table, the rubric applied allows assessment criteria (dimensions) to be linked to attainment levels. At the intersection of the dimensions (columns) and levels of attainment (rows) we find the descriptors, statements that define precisely the features of the dimension described. The rubric contains five possible performance levels for each dimension: Very Good, Good, Acceptable, Below Standard and Unacceptable.

The analysis of some rubrics used in professional translation contexts (for example that of the American Translators Association for certification purposes) has provided an insightful input that helps to outline the quantitative element of the new model. This rubric breaks the concept of translation quality into four dimensions, whose definition derives from the functionalist concept of translation quality based on the notion of *adequacy*. The dimensions refer to the adequacy in the conveyance of the general sense, of the conformance to target language rules and of the general and specialized contents. Fuzzy and blurred as it is, the boundary between general and specialized knowledge is basically established on cognitive terms (Montero, Faber and Buendía 2011). So the task of the rater is restricted to choosing the most appropriate descriptor for each dimension, thus reducing considerably the unavoidable subjective burden of the reviser. Graphically this theoretical model turns into the following figure:

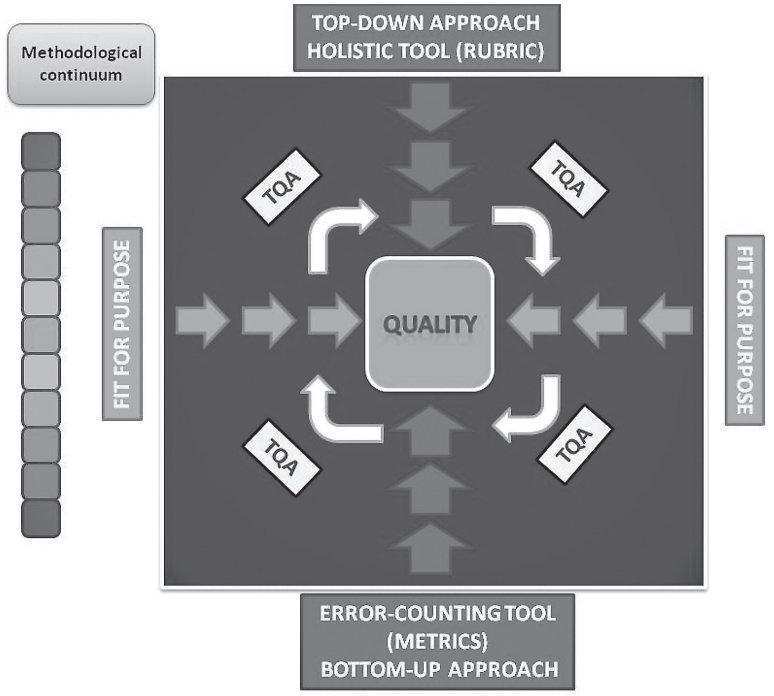

FIGURE 3: Graphical representation of a mixed top-down/bottom-up approach to TQA

As may be observed in the figure, the TQA model puts forward a combined top-down/bottom-up approach by using a quantitative tool (metrics) and an assessment rubric. This mixed approach links the dimensions of the rubric to the error types of the metric and meanwhile the *fitness-for-purpose* principle governs the TQA process. Methodologically, the model integrates two tools from complementary approaches within a *continuum* (Waddington 2000). This flexible tool allows the requester to set the order (of relative importance) of the rubric's dimensions and of the metric's errors by allocating them credit points (rubrics) and deduction points (metrics). Therefore, the resulting tool integrates two complementary views. On the one hand, a top-down approach through the rubric that provides a quantitative assessment of the macrotextual elements of the text by allotting them bonus points. And on the other hand, a bottom-up approach through the metrics that flags and counts error at microtextual level by subtracting points allocated to each error. The application of this componential tool will supply the rater with two quality indicators, one of a qualitative nature (rubric) alongside a quantitative one (metric). Stemming from opposite but complementary views, these two quality indicators will offer the evaluator a global view and a solid basis for making a justifiable decision on the quality of the translation.

This conceptual design remains to be experimentally tested. The results of this continuing empirical study will be disseminated in due time.

Notes

[1]. This paper subscribes the assumption posed by Sir William Thompson in XIX[th] Century: *"You can not assess what you can not measure"* (in Muzii 2006: 21-22) so there is a need to quantify it somehow.

[2]. For example, some general quality definitions identified quality with "fitness for use" or "zero defects" (Juran 1974) or as "conformance to requirements" (Crosby 1979) or as "a system of means to economically produce goods or services which satisfy customers' requirements" (Japanese Industrial Standards Committee 1981).

[3]. For Garvin (LISA 2004: 31) quality is a concept composed of five categories: Perceived; Product-based; User-based; Operations-based and Value-based. These five categories draw a picture whereby quality in translation is a multidimensional reality where each of them adds essential cues to form a comprehensive quality picture; however, none of them on their own would suffice to give a global view of quality.

[4]. TdA: *Sistema Canadiense de Apreciación de la Calidad Lingüística*. Initially created by Alexandre Covacs and afterwards joined by Jean Darbelnet.

[5]. [http://www.translationdirectory.com/article386.htm (Consulted on 7 march 2011)]

[6]. According to LISA (2007: 43), the latest LISA QA Model version (3.1, January 07) is the most widely used tool for TQA in localization and about 20% of all the companies in the world that take part somehow in localized product testing use it. Consulted on 21 March 2011. Available in http://www.lisa.org/LISA-QA-Model-3-1.124.0.html

[7]. Personal insertion

[8]. Emphasis in the original.

[9]. The translation body of the European Commission

[10]. According to DGT's own sources, in 1997 outsourced translations accounted for 16%, whereas in 2004 this figure increased to 23% and in 2008 it reached 26% out of a total translation of 1,805,000 pages. These figures show a clear upward trend in outsourcing percentages and this is expected to continue.

[11]. For further information about the Translation Centre, go to http://cdt.europa.eu/ES/whoweare /Pages/Presentation.aspx

[12]. According to their definition Utility refers to the 'relative importance of the functionality of translated content', Time is the deadline and Sentiment alludes to the 'importance of impact on brand image' (O'Brien 2012: 71)

[13]. 1. User Interface Text, 2. Marketing Material, 3. User Documentation, 4. Website Content, 5. Online Help, 6. Audio/Video Content, 7. Social Media Content, 8. Training Material.

[14]. B2C, B2B and C2C (O'Brien 2011 :68)

[15]. Quality Control is less than a full-revision and Quality Assurance is a broader concept that includes other minor procedures (Mossop 2007: 118)

Works cited

Aenor. 2006. *Norma Española UNE-EN 15038. Servicios de traducción. Requisitos para la prestación del Servicio.* Madrid: AENOR.

Bowker, Lynne. 2000. "Towards a Methodology for Exploiting Specialized Target Language Corpora as Translation Resources". *International Journal of Corpus Linguistics* 5: 17-52.

Brunette, Louise. 2000. "Towards a Terminology for Translation Quality Assessment— A Comparison for TQA Practices". *The Translator* 6: 169-182.

Colina, Sonia. 2008. "Translation Quality Evaluation: Empirical evidence for a functionalist approach". *The Translator* 14: 97-134.

—. 2009. "Further Evidence for a Functionalist Approach to Translation Quality Evaluation". *Target* 21: 235–264.

Conde Ruano, José Tomás. 2008. *Proceso y Resultado de la Evaluación de Traducciones*. (Doctoral Dissertation, Universidad de Granada, Spain). Retrieved from http://digibug.ugr.es/handle/10481/2309 (Accessed 20 March, 2011).

Crosby, Philip Bayard. 1979. *Quality is Free.* New York: McGraw-Hill.

De Rooze, Bart. 2006. *La Traducción, contra Reloj. Consecuencias de la Presión por Falta de Tiempo en el Proceso de Traducción*. (Unpublished Doctoral Dissertation, Universidad de Granada, Spain). Retrieved from: http://isg.urv.es/library/papers/DeRooze-DissDraft03.pdf (Accessed 4 June, 2011)

Doyle, M. S. 2003. "Translation Pedagogy and Assessment: Adopting ATA's Framework for Standard ErrorMarking". *The ATA Chronicle* November/December 2003.

Gómez Gónzalez Jover, Adelina. 2002. "La Equivalencia como Cuestión Central de la Traducción en las Instituciones de la Unión

Europea". In VV. AA. *El español, lengua de traducción. I Congreso Internacional*. Almagro: Comisión Europea y Agencia EFE: 438-457.

Gouadec, Daniel. 1981. "Paramètres de l'Evaluation des Traductions". *Meta* 26: 96-116.

Hönig, Hans G. 1998. "Positions, Power and Practice: Functionalist Approaches and Translation Quality Assessment". In Schäffner C. (ed.) *Translation and Quality*. Clevedon, UK: Multilingual Matters: 6-34.

Horguelin, Paul and Louise Brunette. 1998. *Practique de la Revisión, 3ème Edition Revue et Augmenteé*. Québec: Linguatech éditeur.

House, Juliane. 1997. *Translation Quality Assessment: a Model Revisited*. Tübingen, Germany: Gunter Narr.

Hurtado Albir, Amparo. 2004. *Traducción y Traductología. Introducción a la traductología*. Madrid: Cátedra.

Japanese Industrial Standards Committee (JISC). 1981. Retrieved from http://www.jisc.go.jp/eng/index.html

Jiménez Crespo, Miguel Angel. 2009. "The Evaluation of Pragmatic and Functionalist Aspects in Localization: Towards a Holistic Approach to Quality Assurance". *The Journal of Internationalization and Localization* 1: 60-93.

—. 2011. "A Corpus-based Error Typology: Towards a More Objective Approach to Measuring Quality in Localization". *Perspectives, Studies in Translatology* 19: 315-338.

Juran, Joseph M. 1990. *Jurán y el Liderazgo para la Calidad. Un Manual para Directivos*. México: Edición Díaz de Santos, S.A.

Larose, R. 1998. "Méthodologie de l'Évaluation des Traductions". [A Method for Assessing Translation Quality]. *Meta* 43: 163-86.

Lisa. 2004. *LISA Best Practice Guide: Quality Assurance — The Client Perspective*. Geneva: The Localization Industry Standards Association (LISA).

—. 2007. *The Globalization Industry Primer: An Introduction to Preparing your Business and Products for Success in International Markets*. Geneva: The Localization Industry Standards Association (LISA).

Martínez Melis, Nicole and Amparo, Hurtado Albir. 2001. "Assessment in Translation Studies: Research Needs". *Meta* 46: 272-287.

Martínez, Roberto and Silvia Montero. 2010. "Calidad y Traducción: el Caso de la DGT". *PuntoyComa 118. Boletín de los traductores españoles de las instituciones de la Unión Europea* 118: 3-9.

Montero, Silvia, Pamela Faber and Miriam Buendía. 2011. *Terminología para Traductores e Intérpretes: Una perspectiva integradora. (2nd ed.)* Granada: Ediciones Tragacanto.

Moskal, Barbara M. 2000. "Scoring Rubrics: What, When and How". *Practical Assessment, Research & Evaluation* 7. Retrieved from http://pareonline.net/ getvn.asp?v=7&n=3 (Accessed 12 September, 2012)

Mossop, Brian. 2007. *Revising and Editing for Translators*. Manchester, UK: St. Jerome.

Muzii, Luigi. 2006. "Quality Assessment and Economic Sustainability of Translation". *Rivista Internazionale di Tecnica della Traduzione/International Journal of Translation* 9: 15-38.

Nord, Christiane. 1997. *Translation as a Purposeful Activity*. Manchester, UK: St. Jerome.

—. 2009. "El Funcionalismo en la Enseñanza de la Traducción". *Mutatis Mutandis* 2: 209-243.

O'Brien, Sharon. 2012. "Towards a Dynamic Quality Evaluation Model for Translation". *Jostrans: The Journal of Specialized Translation* 17. Retrieved from http://www.jostrans.org/issue17/art_obrien.php (Accessed 21 October, 2012)

Parra Galiano, Silvia. 2005. *La Revisión de Traducciones en la Traductología: Aproximación a la Práctica de la Revisión en el Ámbito Profesional Mediante el Estudio de Casos y Propuestas de Investigación*. (Doctoral Dissertation, Universidad de Granada, Spain). Retrieved from http://digibug.ugr.es/handle/ 10481/660 (Accessed 5 February, 2010)

Rothe-Neves, Rui. 2002. "Translation Quality Assessment for Research Purposes: an Empirical Approach". *Cuadernos de Tradução* 10: 113-131.

Sae. The Engineering Society for Advancing Mobility Land Sea Air and Space. 2001. *Surface Vehicle Recommended Practice*. Retrieved from http://www.apex-translations.com/documents/sae_j2450.pdf(Accessed 3 July, 2011)

Schäffner, Christina. 1998. "From 'Good' to 'Functionally Appropriate': Assessing Translation Quality". In Schäffner, C. (ed.) *Translation and Quality*. Clevedon, UK: Multilingual Matters: 1-5.

Secâra, Alina. 2005. *Translation Evaluation – a State of the Art Survey. eCoLoRe/MeLLANGE Workshop Proceedings.* Leeds, UK: University of Leeds Press: 39-44.

Stejskai, Jiri. 2006. "Quality Assessment in Translation". *MultiLingual* (June) 80 (17): 41-44.

Waddington, Christopher. 2000. *Estudio Comparativo de Diferentes Métodos de Evaluación de Traducción General (Inglés–Español).* Madrid, Spain: Universidad Pontificia de Comillas.

Williams, *Malcolm*. 1989. "The Assessment of Professional Translation Quality: Creating Credibility out of Chaos". *TTR: Traduction, Terminologie, Redaction* 2: 13-33.

—. 2001. "The Application of Argumentation Theory to Translation Quality Assessment". *Meta* 46: 327-344.

—. 2004. *Translation Quality Assessment: An Argumentation-Centred Approach*. Ottawa, Canada: University of Ottawa Press.

Received: 16 August 2013
Accepted: 6 March 2014