This article explores whether it is possible to synthesise the theorising about translation quality evaluation in the literary translation sphere with the moves in technical translation to establish ‘metrics’ (point-scoring methods) for measuring translation quality. It would be useful, too, to have methods of measurement which can be used for both human and machine translation. That would be a real measure of the quality of machine translation. A customer may well be impressed by claims for machine translation of “90% accuracy”, but in reality that means a mistake in every ten terms, and some of those mistakes may be serious ones.

It is important, first of all, to draw a distinction between, on the one hand, Quality Assurance Certification, which is now widespread in the commercial translation world, and Translation Quality Assessment (henceforth: TQA) itself on the other.

I have identified four basic types of quality assurance procedures in current practice: bottom-up revision, top-down revision, qualifications of performer, and constrained procedures.

Bottom-up revision is where the translation is looked at by a reviser who is more senior to, and more experienced than, the original translator. Four eyes, it is considered, are better than two eyes, and provide assurance that there will be nothing wrong with the translation. Mis-translations, minor mistakes and infelicities will all be corrected. The translator benefits, because with proper feedback he or she benefits from the greater knowledge and competence of the senior translator. The reviser benefits, because a reviser can normally revise several times as fast as producing a translation from scratch.

The big disadvantages are cost, time, resources and the subjective nature of translation. To have two people look at every translation is an expensive luxury which few can afford in today’s harsh commercial environment, where some industrial translation projects can run into millions of words. And it considerably slows the pace at which translations can be produced, and as anyone who has organised translation work knows, the deadlines are often very short indeed. There is no shortage of would-be translators, but there is a serious shortage of experienced translators who are prepared to do revision, a task which offers little job satisfaction. And any competent original translator is going to quarrel with at least some of the revisions, since inevitably there will be those which derive from the subjective preferences of the reviser rather than from objective criteria.

Top-down revision, which is practised in a few translation companies, is where the translator is seen as the expert, possibly with considerable subject expertise, but a less-qualified linguist comes along as a sort of sweeper-up, to make sure nothing has been omitted or figures wrongly transcribed, etc.
‘Qualifications of performer’ is the quality assurance dear to the heart of translator associations. They would all like to require that translations should only be done by qualified professionals.

Constrained procedures is my shorthand term for procedures such as various international and national standards (e.g. ISO 9000, DIN 2345), which can provide parameters governing the manner in which practitioners should approach translation assignments.

All these different types of quality assurance procedures concentrate on the producer or on the process, and have a useful role to play in indicating ways in which better translations may be produced. But none of them measure the product.

And with globalisation, and the exponential growth of translation requirements in the commercial, technical and scientific spheres, and a corresponding increase in the sums spent on translations, a way of measuring the quality of the product is precisely what customers are beginning to demand.

This was exactly what Professor Peter Schmitt of Leipzig University, had in mind when, in an attempt to move the process forward, he organised the first major international conference specifically on translation quality assessment in Leipzig in October 1999.

“There are no generally accepted objective criteria for evaluating the quality both of translations and interpreting performance,” he said. “Even the latest national and international standards for this area - DIN 2345 and the ISO 9000 series - do not regulate the evaluation of translation quality in a particular context. Professional translators rightly expect help from translation theorists, who in turn point to the complexity of the subject matter and who so far have failed to come up with any answers or have suggested criteria and procedures which translators and interpreters find impossible to apply in practice...An industry that is over 2,000 years old and which has an annual turnover of more than 2,000 million marks in Germany alone ought, on the eve of the new millennium, to be developing a clear idea of how to determine the quality of its product, thus creating the basis for practicable quality management systems”.

Fortunately the simplistic attitude of many customers – that ‘a translation is a translation’, and it is either right or it is wrong – seems to be disappearing. It was an attitude which in my early days in organising translations I did come across all too frequently. This applied particularly in the English-speaking countries, which are much less exposed to other languages. To the language ignoramus language transfer is simply a matter of one-for-one substitution.

The average customer does not understand how you can give the same source text to ten different professional translators and have, as a result, ten target texts with a considerable measure of individual variation. He does not understand that translation can only ever be an approximation, that every word in a language is so loaded with nuance and cultural variation that exact equivalence is rarely achieved, even between closely related languages.
It was often argued that while cultural variation was important for literary text, when it comes to technical translations, exact equivalences will be found, because here we are dealing with specific objects. But this is not so. Even here a language’s cultural way of looking at the world, its Weltanschauung, will come into play.

It can be difficult convincing a customer - particularly if he wants an automatic translation of his parts list - that a bolt in English is sometimes a Bolzen in German, but more often than not a Schraube (and, very occasionally, a Stift).

To the German mind they are different parts. Germans see parts in terms of how they are designed, while the English, pragmatic as always, think of their function.

In the same way you have to tell the customer that you cannot translate the English world valve into French unless you know what kind of valve it is - it could be a souppape, a vanne, a clapet, even sometimes a robinet, and to the French mind these are different objects, and the French terms are not interchangeable.

As the above cases exemplify, technical translation (and indeed all translation) often requires supra-textual knowledge, that is knowledge of how things operate, which goes far beyond straightforward linguistic knowledge.

But if translation can only be an approximation on the one hand, and requires supra-textual knowledge on the other, will it ever be possible to establish parameters for objective evaluation of translation quality?

The discussion of what constitutes a ‘good’ translation has been going on for some two thousand years. Early translators came across the same problem as I had with ignorant customers - that the uninformed think that translation is a matter of word-for-word rendering. The Roman orator Cicero (“Non ut interpres sed ut orator”) and the early Bible translator St Jerome (“non verbum e verbo sed sensum exprimere de sensu”) both had to make the case for translating the ‘sense’ and not the ‘word’.

For centuries the arguments raged, usually between those who wanted as close a rendering of the original as possible and those who believed in a ‘natural style’ in the target text. Eventually the ‘natural style’ came to prevail, and is best summed up by Alexander Fraser Tytler in a celebrated book published in 1790: Essay on the Principles of Translation. The importance of this book was recognised by the Dutch publishing Company John Benjamins, and in 1978 they published a facsimile edition¹ of the original publication.

Tytler formulated (page 16) three rules of translation, which have been so influential that it is worth quoting them in full:

“I. That the Translation should give a complete transcript of the ideas of the original work.
Rule I is today reflected in most systems for measuring translation quality - there is usually a category termed “Omission” or “Completeness”.

Rule II encompasses what is today called by the linguistic term “Register” - you do not translate the plain text of an instruction manual intended for dumper truck drivers into more high-flown literary language.

Rule III really did finally lay down the ground rule for “Style” - that you must not let the original language stick through, like a sore thumb, into the translation - German-type embedded clauses in English text, for example.

The first major contribution to formulating a yardstick of assessment came in the 1960s when Dr Eugene Nida\(^2\) introduced the concept of dynamic equivalence, and the related one of equivalent effect. If the target text had the same effect on a target language reader as the source language had on a source language reader, then the translation was successful. And tests, it was suggested, could perhaps be devised to check the level of comprehension by the target reader. The most objective test of all, perhaps, is the commercial one, particularly for sales literature. If an intended readership in a target country is not buying the car, or the photocopier, or the cosmetic product, then however accurate the textual equivalent the dynamic equivalence is not being achieved!

Before Nida the success of a translation was always determined by how accurately it rendered the source text into the target language. By focusing on the effect of the target language text, Nida’s ideas can be said to presage the so-called *Skopos Theory* which was developed in Germany by Katharina Reiss and Hans Vermeer\(^3\). In contrast to the more traditional - “the source text is sacred” - which derives from literary translation, the *skopos* theory argues that the target text has to be established according to the purpose, function or *skopos* which it is intended to fulfil. The *skopos* theory is particularly applicable to technical translation, since such translation is intended to serve the specific purpose of conveying information or instructions.

Juliane House, the name most frequently quoted in the field of translation quality assessment, has re-emphasised the importance of the source text, and indeed the first task in her assessment model\(^4\) is a detailed analysis of the source text. Her approach is based on linguistic stylistics and a Neo-Firthian model. The model analyses the situational dimensions of the source text, and then looks at the translation text to find if there is any “mismatch along the dimensions”. These she calls “covertly erroneous errors”, to be distinguished from “overtly erroneous errors” which result from either mistranslation or from breaches of the norms of the target language. “Cases of breaches of the target language system are subdivided into cases of ungrammaticality, i.e. clear breaches of the language system, and cases of dubious acceptability…”

“Both groups of overtly erroneous errors”, writes House, “have traditionally been given much more attention whereas covertly erroneous errors, which demand a much more qualitative-descriptive, in-
depth analysis, have often been neglected”.

More recently House has emphasised the interaction between source text and target text, and the success of a translation, in her view, is the degree to which it provides a semantically and pragmatically equivalent to the source text.

But House time and time again backs away from any attempt to provide a universally applicable objective basis for translation quality assessment. “Translation evaluation - despite the attempt in my model to objectify the process by providing a set of categories - must consequently also be characterised by a necessarily subjective element, due to the fact of course that human beings are important variables”.

So, it seems, we are no further forward in moving translation quality assessment from the subjective sphere to the objective sphere.

The concept of metrics was introduced into the debate at the 1999 Leipzig conference mentioned above, with a paper by Dr Kurt Godden describing the J2450 project of the Society of Automotive Engineers (SAE), a project which since then become an SAE standard, and has been widely adopted by both US and European automotive companies and by their major translation suppliers.

The concept of metrics was not new. They had been used in comprehensibility rating for informative texts, and in technical writing, for many years, and also in quality assurance systems such as Six Sigma. Six Sigma, which originated in Motorola but was taken up by other companies, started out as a ‘constrained procedures’ quality assurance system, but a requirement quickly developed for ways of measuring design or production performance. And on a small scale every university where marks were given to translation examination papers must have tried to have reasonably objective criteria, however approximate. But J2450 was the first major attempt to apply metrics in translation on a large scale.

That J2450 metric was designed originally specifically for automotive service literature, which is one reason why ‘style’ is ignored. Basically there are seven error categories, and every error is evaluated as either serious or minor. The categories, with the number of points which are given to error, are as follows.

<table>
<thead>
<tr>
<th>Error category</th>
<th>Serious</th>
<th>Minor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wrong Term</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>Syntactic Error</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Omission</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Word structure or agreement</td>
<td>4</td>
<td>2</td>
</tr>
</tbody>
</table>
One surprising feature of the metric is that it did not have a category for mis-translation (where the translator has misunderstood something in the original). The designers of J2450 were confident that the problem of mis-translation could be covered by the category “Wrong Term”, but this has continued to be a source of considerable debate at the many meetings which have been held by users.

Another feature which sometimes raises eyebrows is that errors in the translation which are caused because the source text was itself in error are not forgiven - they are still regarded as errors. This really dramatically shifts the emphasis to the target text. A translator is no longer a passive conduit, but a technical writer with full responsibility for the text he or she produces.

An echo of this can be found in the writings of a translation scholar, Malcolm Williams, who suggests that we should adopt what he calls an argumentation-centred approach to translation quality assessment, whereby the important criterion is to what extent the translation conveys the ‘argument’, or core message, of the original. I noted in particular one phrase:

“Thus the mistranslation of an individual word, phrase or sentence in the translation is not analysed from the standpoint of degree of equivalence to the corresponding words in the source, it is judged according to the contribution that the source text makes to the purpose, or elocutionary point, of the text…”

Reverting to J2450, the lack of any room for manoeuvre in how many points are awarded for an error once the decision has been made on major/minor is not felt to be a problem in practice. “While sometimes this assignment of numeric weights will over-value the severity of an error, it will under-value it at other times. The underlying assumption of SAE J2450 is that these deviations will tend to cancel each other…”

Because J2450 is a measurement tool rather than a revision tool, it does not address what to do in the event of a Disaster Error (sometimes called Critical Error), an error so serious that it could be life-threatening or otherwise highly injurious to the customer’s interest if it were allowed to go uncorrected.

The Localization Industries Standards Association (LISA) produced its own metric for assessing the quality of software localisation, which basically sets out parameters for assessing translation quality, and is less prescriptive than J2450. The translation company ITR in the UK brought out a software program
called Blackjack which can be used as a semi-automatised tool for assessing TQA. Blackjack uses 21 error categories (the first of which, it is interesting to note, and in contrast to J2450, is ‘misinterpretation of source language text’) and a scoring system of 0 to six for each error.

With 21 error categories Blackjack is obviously a more complex tool to implement than J2450, and it does allow for feedback on errors with a view to remedial action.

The largest translation organisation in the world, the Translation Service of the European Commission, has been steadily increasing the amount of work it puts out to freelance suppliers. It therefore needs a reliable TQA system, and so evolved what is called the External Translation Evaluation Interinstitutional Procedure. Again, various parameters are used, including concerns particular to the European institutions. These include whether reference material supplied was used (consistency of style and terminology is essential for EC documentation), and whether the translation was delivered on time (a perfect translation three days after the meeting has taken place, whatever its other qualities, is worthless). Mistranslation is a major category, and can be marked as “high relevance” or “low relevance”. A high relevance error is one which seriously compromises the translation’s usability, and therefore similar to “critical error” or my own term “disaster error”.

What conclusions can we draw?

The first conclusion must be that metrics are here to stay. Customers (normally) are not going to revise the translations themselves, so they need to have some assurance that what they are getting is of satisfactory quality. Before-the-event quality assurance procedures, such as practitioner-qualification and constrained-procedures, can go some way to providing that assurance, but they need to be supplemented by a reasonably objective after-the-event way of measuring translation quality.

The second conclusion which has emerged from experience is that the identification of error categories and the weighting they are given depend very much on the type of text involved. My own opinion is that more work needs to be done on text typology, but that most large-scale translation projects do fall into a reasonably small number of easily-definable text types.

The experience, particularly with J2450, shows that the establishment of TQA processes is lengthy and time-consuming. A third conclusion therefore is that there should be no unnecessary re-inventing of the wheel. Experiences need to be pooled and shared.

A fourth conclusion, emerging from J2450 experience, is the importance of training (and evaluating) the evaluators, so that everyone is singing from the same hymn sheet. The maverick evaluator, who categorises even the tiniest punctuation error with the severest mark possible, or the over-indulgent linguist who fails to imagine the consequences of an error, can wreck confidence in the procedure. We can learn a lot here from the experience of language (acquisition) testing, which has become something of an exact science. One day there could be an Institute of Translation Quality Assessment, perhaps located in one of the interested universities.
A fifth conclusion is to recognise that translations are not produced in isolation, in unreal laboratory conditions. The European Commission’s inclusion of reliability in delivery as a factor in evaluating TQA must surely be taken up in other systems because of the importance of the issue.

A sixth conclusion is that while the main lines of the solution are now emerging - error categories, numerical weighting of errors, adjustments according to text type etc. - there is still room for discussion on the detail, and here those who work in the industrial field should be trying to involve translation studies academics. Can macro-content or situational dimensions be measured? Do we need to bring in comprehensibility ratings (Eugene Nida was making some interesting suggestions in this regard 30 years ago) to ensure that the message is being effectively communicated?

And a seventh conclusion must be that we must remain constantly aware of wider implications. Can TQA be paralleled with similar procedures for evaluating the quality of source texts? How do the TQA procedures we are evolving fit into systems such as Six Sigma? Do we need to import more expertise from the field of statistical sampling? Can we extend the principles from text to other media - images or sound, for example (using images of alcohol or scantily-dressed ladies in material for Arabic countries would be definitely a ‘wrong term’ type of serious error!).

The approach which I suggest, is one I call PEXIS (Purpose-oriented Explicit or Implicit Specification). My argument is that you can only determine the quality of a translation when you know the specification. Often there will not be an explicit specification (though DIN 2345 suggests there should be) but in many cases the text type and intended readership would suggest an implicit specification. For example, if you are translating a patent application for filing purposes, there is a particular set of parameters which must be followed; different criteria might apply if the translation is for a customer who basically wants to know what his competitors are up to. In a Bible translation, the translator needs to know whether he is translating the original in order to provide a new exegesis for scholars, an easily comprehensible ‘vernacular’ version for those who want to know what the Bible tells us, or a ‘sacred’ version which uses high-flown language and will often require the use of sonorous phrases familiar from previous translations and which are part of a heritage.

Reiss and Vermeer did anticipate to some extent the specification approach when they asserted that the skopos or purpose of a translation “can be said to vary according to the recipient”.

And resulting from this it seems to me that we need much more work on text typology, and subsequently on what I call ‘implicit specification’, so that we can develop more explicit parameters for defined text types.

This is normally already the case for technical writing. As Susanne Göpferich has put it:

“When technical writers produce texts, they are not normally doing this to satisfy their own personal
communication needs, and to discuss their personal findings, but in response to an assignment which consists of creating a text with a specific function which has been specified by the customer.” (My English translation of the German original).

[“Wenn Technische Redakteure Texte einstellen, tun sie das in der Regel nicht, um einem eigenen, persönlichen Kommunikationsbedürfnis nachzukommen und eigene persönliche Erkenntnisse zu versprachlichen, sondern um einen Auftrag zu erfüllen, der in der Anfertigung eines Textes mit einer vom Auftraggeber spezifizierten Funktion besteht.” (Susanne Göpferich)]

But what is also required is a study of how technical texts are produced *ab initio*. We may well then discover that there are important cultural differences. I have never had the opportunity of researching this, but my anecdotal impression is that, for example, German readers like to have an overall structure. Therefore, for example, instructions for assembling and operating a lawnmower in German would begin with a situational introduction, of where and when the lawnmower should be used. English readers are more pragmatic and would look first of all for the assembly instructions. It may be argued, therefore, that a translation from German to English might well omit the Introduction, and conversely a translation from English to German might insert such an introduction.

But this is by no means an original view. It was put forward over 20 years ago by the Finnish scholar Justa Holz-Mänttäri. I do not have a copy of her work available to me, so I quote Cay Dollerup’s summary of one of her principal arguments.

“She emphasises the primacy of establishing a target text which will function adequately in the target culture in specific contexts. She downplays the source text (and consequently a source- and target-text comparison). In her view, the source text is only one element in the source material. The source material derives from the complete social, indeed cultural, context in which the source text exists. This material comprises all elements in the chain of translational communication, from the writer(s) of the original, the initiators, specialists, etc. It also includes the target-text recipients, and the translator’s role is first and foremost to act as the intercultural expert who can tell how the target text should be phrased in order to fulfil its function in the target-language situation.”

But this sort of thinking could lead us to explore the interface between translation and technical and scientific writing which could provide enough material for a quite separate article, perhaps with the title *Text Generation from Source Material in another Language*!
References:

7. ITR: www.itr.co.uk.